(Vision Paper) A Vision for Spatio-Causal Situation Awareness, Forecasting, and Planning*

FAHIM TASNEEMA AZAD†, K. SELÇUK CANDAN, AHMET KAPKIÇ, MAO-LIN LI, HUAN LIU, PRATANU MANDAL, and PARAS SHETH, Arizona State University, USA

BILGEHAN ARSLAN‡, Arizona State University, USA

GERARDO CHOWELL-PUENTE, Georgia State University, USA

JOHN SABO, Tulane University, USA

REBECCA MUENICH, University of Arkansas, USA

JAVIER REDONDO ANTON and MARIA LUISA SAPINO, University of Torino, Italy

Successfully tackling many urgent challenges in socio-economically critical domains, such as public health and sustainability, requires a deeper understanding of causal relationships and interactions among a diverse spectrum of spatio-temporally distributed entities. In these applications, the ability to leverage spatio-temporal data to obtain causally-based situational awareness and to develop informed forecasts to provide resilience at different scales is critical. While the promise of a causally-grounded approach to these challenges is apparent, the core data technologies needed to achieve these are in the early stages and lack a framework to help realize their potential. In this paper, we argue that there is an urgent need for a novel paradigm of spatio-causal research built on computational advances in, spatio-temporal data and model integration, causal learning and discovery, large scale data- and model-driven simulations, emulations, and forecasting, spatio-temporal data-driven and model centric operational recommendations, and effective causally-driven visualization and explanation. We, thus, provide a vision, and a road-map, for spatio-causal situation awareness, forecasting, and planning.

Additional Key Words and Phrases: spatial algorithms, spatial bigdata, causal discovery

*This research has been funded by NSF#1909555 "pCAR: Discovering and Leveraging Plausibly Causal (p-causal) Relationships to Understand Complex Dynamic Systems", NSF#2311716 "CausalBench: A Cyberinfrastructure for Causal-Learning Benchmarking for Efficacy, Reproducibility, and Scientific Collaboration", NSF#2125246 "PanCommunity: Leveraging Data and Models for Understanding and Improving Community Response in Pandemics", NSF#2200140 "Predicting Emergence in Multidisciplinary Pandemic Tipping-points (PREEMPT)", US Army Corps of Engineers Engineering With Nature Initiative through Cooperative Ecosystem Studies Unit Agreement #W912HZ-21-2-0040, and EU#955708 "EvoGamesPlus: Evolutionary Game Theory and Population Dynamics".

†The order of the authors are alphabetical by last name within each institution and does not reflect their contributions to the manuscript.

‡The work was done when the author was at Arizona State University. The author's current affiliation is Gazi University, Turkey.

Authors' addresses: Fahim Tasneema Azad, fazad@asu.edu; K. Selçuk Candan, candan@asu.edu; Ahmet Kapkiç, akapkic@asu.edu; Mao-Lin Li, Mao-Lin.Li@asu.edu; Huan Liu, huan.liu@asu.edu; Pratanu Mandal, pmandal5@asu.edu; Paras Sheth, psheth5@asu.edu, Arizona State University, USA; Bilgehan Arslan, Arizona State University, USA, barslan1@asu.edu; Gerardo Chowell-Puente, Georgia State University, USA, gchowell@gsu.edu; John Sabo, Tulane University, USA, jsabo1@tulane.edu; Rebecca Muenich, University of Arkansas, USA, rlogsdo@uark.edu; Javier Redondo Anton, javier.redondoanton@unito.it; Maria Luisa Sapino, mlsapino@di.unito.it, University of Torino, Italy.

Manuscript submitted to ACM

ACM Reference Format:
Fahim Tasneema Azad, K. Selçuk Candan, Ahmet Kapkiç, Mao-Lin Li, Huan Liu, Pratanu Mandal, Paras Sheth, Bilgehan Arslan, Gerardo Chowell-Puente, John Sabo, Rebecca Muenich, Javier Redondo Anton, and Maria Luisa Sapino. 2024. (Vision Paper) A Vision for Spatio-Causal Situation Awareness, Forecasting, and Planning. 1, 1 (September 2024), 41 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Data technologies have the potential to completely transform how we address global issues by providing insights that could help us make better, more informed decisions in fields like environmental sustainability and public health, which are vital to our future as a society. This promise is not without challenges, requiring an examination of the causality and the dynamic interconnections that exist throughout our natural and socioeconomic systems:

- **(Water Sustainability)** Natural infrastructure tends to be built in a haphazard, yet highly interconnected, way. Natural water infrastructure, for example, can include (a) a network of wetlands, (b) healthy soils, (c) forest ecosystems, and (d) snowpack, and this infrastructure provides services that include flood protection, erosion control, water storage, and purification. Built water infrastructure that complements natural water infrastructures comprises two entire sectors: (e) the dams sector and (f) the water and wastewater systems sector. Management of this causally-interlinked, yet spatio-temporally distributed, water infrastructure is a pressing contemporary challenge given a changing climate and increasing water demand for agriculture and human consumption [4, 21, 36, 103, 118, 125].
- **(Epidemics and Pandemics)** The emergence and propagation of an epidemic involves a causally complex interplay of spatio-temporally distributed entities in a multi-layer network, including (a) spatially distributed individuals and their social-interactions, (b) physical short-range and long-range networks of mobility, (c) parameters of disease models (such as infection rate, average length of recovery, and impact of treatment), and (d) intervention decisions (such as school closures or restrictions on mobility) by independent decision makers (city, state, national governments and international agencies, as well as various local and global corporations) [9, 10, 22, 27, 39, 47, 51, 65, 89, 112, 117, 127, 176–178].

Both human decision-making and model discovery in such applications can be significantly improved if these data sets are analyzed for key causal features to discover the underlying latent structures and spatio-temporal dependencies critical for modeling, situation-awareness, forecasting, optimization, and control. In addition, causal learning may take advantage of physically based models' advantages, and causal knowledge may help make data-driven learning more generalizable across different but similar spatio-temporal contexts. Yet, while the promise of a causally-grounded approach to these challenges is apparent, the core data technologies needed to achieve these are in the early stages and lack a framework to help realize their potential.

1.1 Challenges – A Case for "Spatio-Causal" Research

As detailed in our report titled "Data Integration in the Service of Synthetic Research" [88], research of the scale that grand challenges require, demands for the ability to explore complex sets of primary data and models to discover important cross-dataset and cross-community patterns that could never be seen when comparing higher-level interpretations. Indeed, natural and built systems consist of *complex physical processes operating at varying spatial and temporal scales*. They involve heterogeneous multi-modal (temporal, spatial, networked) data and models, 100s of inter-dependent parameters, spanning multiple layers and geo-spatial frames, affected by complex dynamic processes operating at different scales and resolutions. Many of these dynamically evolve due to how the underlying systems and ecosystems develop and due to the preventive and reactive actions taken by individuals and public interventions, requiring continuous adaptation and remodeling. Note that, in addition to their benefits, the available intervention options (lockdowns, travel controls, vaccination programs) can have significant negative socioeconomic impacts, raise health concerns, and may require behavioral changes incompatible with a population's belief systems or may allocate limited resources unevenly across populations, which may be competing for those resources. These increase the importance of situational awareness and increase pressure to conduct research that supports fair, bias-free, and explainable predictions and decisions. In particular, given the *unpredictability of the natural and human factors* and the *complex trade-offs among system-level and human-centered objectives*, decision makers often need to generate many thousands of simulations, each with different parameters corresponding to different, but plausible scenarios. Running and interpreting such spatio-temporal simulation results to generate timely actionable results are computationally costly and, consequently, data and simulations are inherently sparse.

Given these, in this paper, we argue that to effectively support critical decision-making in complex and dynamic human-centered environments, a novel paradigm of **spatio-causal research**, built on computational advances in spatio-temporal data and model integration, causal learning and discovery, large-scale data- and model-driven simulations, emulations, and forecasting, spatio-temporal data-driven, and model-centric operational recommendations, and causally-driven visualization and explanation are needed. For example, detecting and correcting inefficiencies in a water system may require causal impact analysis and fault prioritizing/ranking that consider complex causal relationships among spatio-temporally networked natural and built entities. Similarly, in emergency response and pandemic scenarios, we need spatio-temporal models that can generalize across time and space among many variables in changing and complex situations through causal discovery. In order to tackle these application challenges, the proposed **spatio-causal research** paradigm must synergistically integrate data-driven and physically-based techniques in the presence of the following core data challenges:

- (Challenge 1 – spatio-temporal complexity and heterogeneity) Decision makers often face multi-scale, multi-layer, and heterogeneous spatio-temporal processes and spatio-temporally distributed observations and data. However, effective decision-making involves more than integrating heterogeneous data and models. It involves taking into account software and human agents with different (and possibly contradictory) objectives/reward functions.
- (Challenge 2 - Dynamicity and context-sensitivity) The world is not static: a disease evolves, and interventions (e.g., school closures) change the way humans interact in space. Causality may be context-dependent, and the contexts themselves may be dynamic and/or have emergent properties of the spatio-temporal system. Moreover, two or more entities in one spatio-temporal context may interact collectively to impact other entities, possibly in other spatio-temporal contexts.
- (Challenge 3 - Computational challenges) The above challenges are complicated by the fact that situation awareness, forecasting, and planning need to address the computational challenges arising from the need to acquire, clean, analyze, index, and search, in a scalable manner, large volumes of spatiotemporal, multi-variate, multilayer, multi-resolution, interconnected, and often incomplete/imprecise data. Causal analysis algorithms themselves are computationally expensive and fine-grain spatio-temporal systems may be expensive to causally analyze. Moreover, traditional causal discovery algorithms may confront issues, such as unobserved variables, and often rely on strong assumptions, such as faithfulness [170], which may need re-interpreted in the spatio-temporal context, taking into accont the spatio-temporal structure of the complex system.

- **Service #1**: integrated, federated, and scalable data and model storage, analysis, and simulaiton
- **Service #2**: making unstructured data queriable, prioritize and rank data, correlate and identify the gaps in the data
- **Service #3**: spatio-temporal and network analytics
- **Service #4**: model discovery, causal discovery, including new emerging patterns
- **Service #5**: going back in history to validate models and going forward into future to support forecasting and if-then hypothesis testing.
- **Service #6:** multi-dimensional optimization of planning & operations
- **Service #7:** visualization to demonstrate multi-objective targets and project ROI

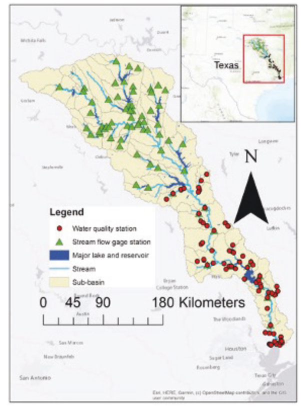


Fig. 1. Key data services for supporting water sustainability (US Army Corps of Engineers Engineering With Nature Initiative, Cooperative Ecosystem Studies#W912HZ-21-2-0040): spatio-temporal causal knowledge is critical in effectively delivering these services

1.2 Outline of this Vision Paper

In this paper, we provide a vision and a road-map for spatio-causal awareness algorithms and applications. In Section 2, we provide an overview of our two motivating applications: (a) hydrology and water sustainability and (b) epidemics and pandemics. We follow this with a brief discussion of the critical concepts in the causality of Section 3 and, subsequently, in Section 4, we introduce the components of a novel framework for describing spatio-temporal causal knowledge. We then discuss research directions toward spatio-temporal causal learning (including models for discovery and inference) in Section 5. In Section 6, we outline how spatio-temporal causal knowledge can be leveraged for effective spatial data structures, models, and algorithms. We conclude the paper in Section 7.

2 MOTIVATING APPLICATIONS

In this paper, we consider two motivating applications, water sustainability and pandemics/epidemics, because these applications urgently need advances in spatial-causal research. In particular, we highlight the need for organically-integrated support for causality in spatio-temporal learning and inference and for effectively considering spatio-temporal relationships in causality learning in these two representative application domains.

2.1 Motivating Application #1: Hydrology and Water Sustainability

As we discussed in the introduction, managing water infrastructure is a pressing contemporary challenge given the changing climate and increasing water demand for agriculture and human consumption. Computational models have the potential to inform the operation of built and natural infrastructure to increase the resiliency of a water infrastructure network and support the design of operational plans to mitigate the negative impacts of adverse events (Figure 1). To achieve this potential, however, underlying models integrating spatio-temporally distributed yet causally-interlinked components, including levees, dams, reservoirs, wetlands, and agricultural lands upstream or adjacent to these, among other infrastructure, must provide accurate operational plans and recommend holistic mitigation actions against droughts and floods.

Wetlands and watersheds are critical natural assets to mitigate the effects of climate change [45]. Optimal location and restoration of wetland infrastructure is a high-dimensional, complex multi-objective problem. Moreover, water quality is a spatio-temporally complex outcome, with multiple causal drivers and potential control points, beginning with the sources in uplands, transport across the landscape and stream networks, and finally, reservoirs. Conventional physically based models have long yielded promising results, as they have been the primary tool to depict the underpinnings of the physics governing the hydrological events. While physically based models can help, applications of these physically based models to operational decision-making have been limited by substantial efforts needed to set up and calibrate the models for a given study area and high computational expense. Causal learning, including deep learning-based methods, is well suited to develop fast surrogates to enable decision-speed analytics [37, 38, 63, 90, 103, 106, 107, 173, 174, 187]. In addition, when trained using data representing a wide range of variability, causally-informed models can be generalizable and transferable to new locations and scenarios [120]. As exemplified with sample results presented in Section 5, the development of native spatio-causal models, combined with rapid increases in computational abilities (graphics processing units, computer clusters. etc.), has the potential to enable hydrologists to utilize data-driven models in tandem with the well-established hydrological models to simulate miscellaneous environmental processes nimbly, and therefore circumvent the conundrums associated with the purely data-driven and purely physically based models. More specifically, intermediate variables of physically based models (e.g., hydrological variables) can train machine learning models that are interpretable and robust against spurious correlations via spatio-causal research that answers the following questions: (1) what types of intermediate spatial variables should be intervened such that we can avoid spurious correlations, and (2) what proper values or distributions would be used to create the interventions? We can augment the original data to train interpretable and causally-robust spatio-temporal models with interventional data from the physically based models.

2.2 Motivating Application #2: Epidemics and Pandemics

Today, experts agree that transformative impact in the effective management of future pandemic threats will arise only through information-driven planning and decision-making for prevention, preparedness, response, and recovery [179]. Casualties and damages could be substantially decreased by proactively responding to an epidemic before (e.g., by pre-positioning response units and supply and service systems), during (e.g., by managing the scheduling and delivery of essential supplies and services while minimizing the unintended consequences of recommended responses), and after (e.g., by the quick implementation of supply-chain networks) the epidemic [62, 150, 151].

Epidemics are the result of complex interplay of spatio-temporal processes (Figures 3). The emergence of novel zoonotic diseases over the last century is primarily a function of environmental and land use changes that affect spatio-temporal interplay among humans, vectors, and wildlife reservoirs [40, 85, 92, 93, 110, 152, 198]. Spread, furthermore, is determined by the spatio-temporally based socio-economically constrained networks that bring susceptible and infected individuals together, including social contact, trade, supply, and transportation networks [68, 126] – human and goods travel provides a pathway for disease movement; yet, travel patterns themselves can be significantly affected by epidemic dynamics, and changes in these patterns can indicate other behavioral shifts that are critical in understanding the epidemic evolution. In short, the emergence and propagation of an epidemic involves a causally complex interplay of entities in a multi-layer network, including (a) individuals and their social interactions, (b) physical short-range and long-range networks of mobility, (c) parameters of disease models (such as infection rate, average length of recovery, and impact of treatment), and (d) intervention decisions (such as school closures or restrictions on mobility) by independent decision makers (e.g., city, state, national governments). Consequently, a better understanding of the networks over

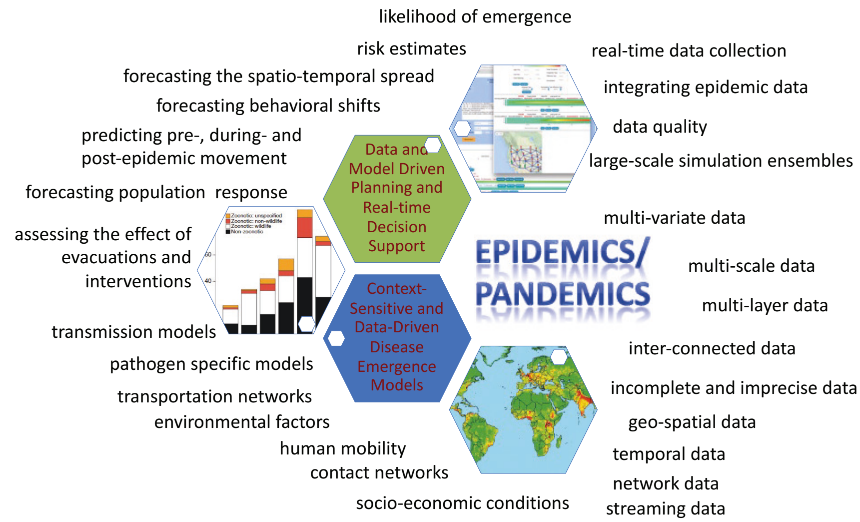


Fig. 2. Elements of data- and model-driven approaches to planning and response to epidemics and pandemics (NSF2125246 "PanCommunity: Leveraging Data and Models for Understanding and Improving Community Response in Pandemics", NSF2200140 "Predicting Emergence in Multidisciplinary Pandemic Tipping-points (PREEMPT)"): spatio-temporal causal knowledge is critical in effectively leveraging data for effective decision making

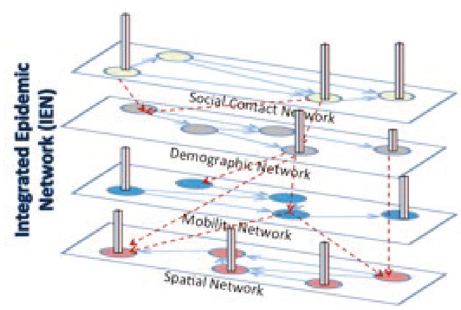
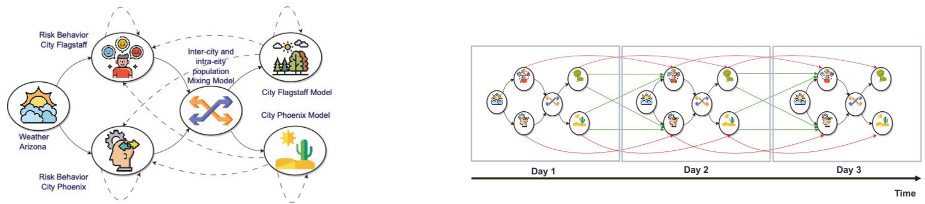


Fig. 3. Multi-layered spatio-temporal epidemic data (interconnections between network layers are denoted with dotted lines; time series corresponding to the various network nodes are denoted with bars)

which disease spread occurs has the potential to improve our ability to generate epidemic predictions [31, 186] and data-driven large-scale computational models of infectious disease emergence and spread are increasingly becoming part of the toolkit to prepare for and manage health emergencies through pharmaceutical and non-pharmaceutical control measures. These include (a) improving our ability to forecast infectious disease emergence and identifying root causes and drivers of disease outbreaks in any region along with community-specific disease transmission vectors, (b) capturing emergent properties and spread characteristics of epidemics [2, 6, 41, 122] within particular network structures derived from network models and real-world networks [2, 6, 41, 122, 175], (c) forecasting the effects of interventions [25, 52, 55, 64, 194], and (d) identifying the effects of private responses to evolving disease threats [49, 50]. The potential for pandemics to rapidly generate mortality [35, 82] and economic impact [84, 123, 146, 169] highlights



(a) Spatio-causal interplay among various factors (b) Causal impact traces over time

Fig. 4. Spatio-causal interplay among various factors impacting pandemic evolution over time

the need to develop new computational frameworks for supporting the science of epidemics and public health, that take into account geospatial context- and network-coupled, dynamic disease emergence and spread models, including evolving transmission patterns, local social and demographic variables, and evolving intervention strategies (including cost and constraints on the interventions) and their side effects. Unfortunately, silo-based modeling, where disease, population dynamics, transportation, and resource models are not integrated, fail to provide an end-to-end view with useful causally-informed predictions for the changing situations, infrastructure conditions, and demands [9, 13, 14]. Existing solutions often rely on hard-coded models, which makes it difficult to integrate new data and models. What is needed instead is the ability to causally stitch together multiple, potentially independently developed, off-the-shelf, component models/simulators, each capturing a different natural, human, or built component of the epidemic in space and time.

To summarize, achieving the aforementioned integration of the multitude aspects of epidemics and pandemics, for predicting their emergence and spatio-temporal dynamics and for planning effective intervention and response strategies in space and time, requires spatio-causal research that natively captures the underlying spatio-causal interplay among the disease, humans agents (carriers, patients, and decision makers), mobility networks, and disaster response networks (Figure 4). While a dynamic multiplex network (mathematically a tensor) can naturally represent the epidemiological data (including spatial network, mobility network, demographic network, and social contact network) to help capture the inherent dynamic interactions between different layers, identifying high-risk groups and developing targeted interventions to control the spread of the disease, requires the ability to trace spatio-causal diffusion patterns and disaggregate various cause-effect structures [161].

3 PRELIMINARIES

3.1 Cause and Effect

Causality is a relationship between an effect and the cause that gives rise to it and causal effect is defined by the strength of the causal relationship (e.g. *"if she gets in contact with a person infected with COVID-19, she will get infected"*). Study of causality has a long history, yet defining what a causal relationship (let alone discovering causal relationships from data) has been a challenge that has not yet been satisfactorily achieved [69]. Most of the early attempts have been statistical in nature. Commonly used Granger causality [8, 104], for example, is statistical in nature. Fisher [54] and followers advocated a statistical approach to causality, rooted in randomized controlled trials (or at least quasi-randomized experiments [34]) to eliminate confounders in the data. Example applications of these include regression discontinuity design (as in [5, 24]) and instrumental variable methods (such as those used in estimating the causal effect of social contagion on exercise [7] and that of policies on economics and political outcomes [44]). Yet, these

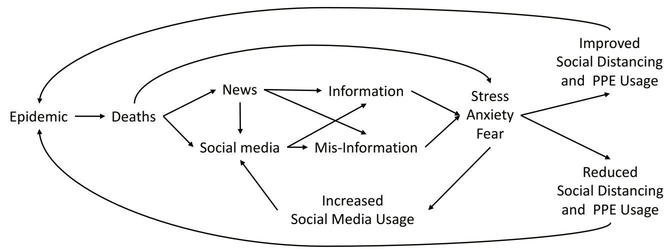
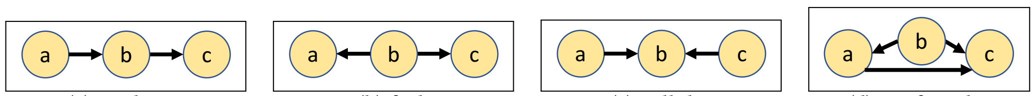


Fig. 5. A sample (non-spatial) causal graph



(a) mediator (b) fork (c) collider (d) confounder

Fig. 6. Basic causal structures: Mediator, fork, collider, and confounder sub-structures describe how causes and effects interact and how causal impact diffuses within a causal system; causal graph based approaches to decision support leverages these local sub-structures as well as global properties (such as d-separation) of the causal graph to determine causal effect – a shortcoming of the existing models for causal representation is that they do not take into account the overlaying of causal and spatial diffusion in spatial systems

approaches rely on certain strong assumptions, which result in paradoxical outcomes when violated. Rubin popularized the idea of "potential outcomes" and the use of counterfactuals in the definition of causality [153]; and deemed causal inference as a missing-data problem and imputation as a valid approach to tackle this challenge. While this led to major breakthroughs in data-driven causal investigations, including the structural equation models [76, 131], its error-free application relies on the applicability of an "ignorability" assumption, which implies that the underlying process is free of unobserved confounders [132–134]. Many studies rely on this strong assumption [58–61, 97]. Unfortunately, strong ignorability and various other commonly used assumptions are not testable and may not be satisfied in the vast majority of datasets [75, 108, 115, 132–134, 142, 145, 149, 192]. In a recent work [147], for example, we proposed to replace the strong ignorability assumption with weaker and testable assumptions which are learned from data.

Unlike these statistics based approaches, others, such as Wright [193], argued that one cannot draw causal conclusions from data, without some causal hypotheses. This line of thought led itself to very successful practical approaches, including path analysis [43, 56, 193], structural equation modeling (SEM [76, 131]), and Bayesian Networks [128], each of which representing contextual knowledge in the form of directed graphs. In its most basic form, a causal graph $G = (V, E)$ is a directed graph that describes the causal effects between variables, where V is the node set and E the edge set. In a causal graph, each node represents a random variable (including the treatment, the outcome, and other observed or unobserved variables in the application domain) and a directed edge $x \longrightarrow y$ denotes a causal effect of x on y. For example, a causal graph in the context of epidemics and pandemics, can describe that deaths from epidemic reported in media and social-media can contribute to awareness as well as fear and anxiety, which may have positive or negative impact on social distancing and PPE usage, which may in turn impact the evolution of the epidemic (Figure 5). More recently, Pearl introduced structural causal models (SCMs [130, 139]) that use directed graphs to explicitly capture causal background knowledge to be enforced or causal hypothesis to be verified or rejected. Pearl and his colleagues have shown that analysis of causal graphs in terms of the underlying causal structures (Figure 6) can be used to avoid many of the common fallacies faced by most purely statistical approaches to causal analysis and that these graph-based techniques offer a principled solution to the treatment of colliders, confounders, and other sources of faulty causal reasoning [131, 132]. Colliders, for example, are related to the explaining away effect [86] and, in epidemiology, they are associated with a type of selection bias (also known as the Berkson's paradox [16]). Confounders also result in many paradoxical scenarios, including the well-known Simpson's paradox [168, 172]. Common applications of this causal graph based approach include the d-separation property, which can be used to determine which independencies should hold in the data [19, 66, 73, 129, 137, 138, 185], front-door and back-door adjustments against confounders [67], c-decomposition technique for finding minimal separating sets [180, 181], and the do-calculus for calculating the effect of an intervention [167]. These have been successfully leveraged for tackling various data challenges including sampling bias [11] and missing data imputation [114].

3.2 Alternative Models for Representing Causal Knowledge

Over the years, there have been various attempts to capture and represent causal knowledge [19, 71, 133, 144]. In this section, we briefly discuss some of the major alternative causal models and if and how they represent the key causal properties, such as causal strength and time lag.

Neuron diagrams: Neuron diagrams [71, 72, 95] represent a causal relationship using a set of neurons with two phases, active and inactive; connections between neurons can be inhibitive or stimulative. Neurons provide a basis for causality that allows some of the plural cause-effect relationships. However, their plural cause structure does not distinguish between AND or OR relationships between two causes. Moreover, they do not represent time and, consequently, self-causation.

Boolean functions: A series of logic gates and variables may also be used to model a cause-effect relationship graph. Similar to neuron diagrams, each cause A may take 0 (absent) or 1 (present) values and these may influence (1) the effect B, or not (0). While the Boolean function model answers the plural-cause ambiguity seen in neuron based representations, this formalism also does not account for time and limits the interactions between causes-effects to Boolean values, ignoring varying strengths of causes [17].

Petri nets: Petri nets are dynamic modeling systems [144]. A simple Petri net structure is defined by $N = (P, T, F)$, with a set of places P (commonly represented as circles), transitions T (represented as rectangles), and flow relations F between places and transitions (represented by arcs), sometimes divided into two separate functions (I, O), input functions I that define arcs from a place to a transition, and output functions O that define arcs from a transition to its destination places. Places represent both causes and effects. However, unlike the other alternative models, places are not directly connected by arcs; instead there is a transition between every place-place connection.

Transitions allow more than single input or output arcs, which allows concurrency of several causes or effects. Arcs can be inhibitive or stimulative. Petri nets share some characteristics with neuron diagrams; however, Petri nets allow extensions, increasing their applicative range. Colored Petri nets [81], for example, introduce classes and the ability to assign specific values to tokens; timed Petri nets allow timed transitions [202], allowing for causal relationships with different lags. The original Petri net model supports plural-causes/effects or concurrency, strength, and feedback properties of a causal system, and its extensions allow richer causal structures.

Structural equation models: Introduced by Wright [193], Structural Equation Modeling (SEM) refers to a large set of methods to represent relationships among variables in a system. Within the context of causal models, SEMs are approached by several authors including Hall and Pearl [19, 71]. These are similar to Petri Nets in terms of flexibility and adaptability as they leverage both latent and known variables and allow differentiating between weak and strong causal assumptions. However, this adaptability also brings ambiguity as the representation (strong causal assumptions require rationale [20]) and formulation of the relationships and assumptions often are not based on well-defined principles.

Bayesian networks: Bayesian networks [136] provide a probability-based representation of a set of causal relationships. Differently from Boolean/binary values, these utilize Bayesian inference to associate a probability to an effect happening given the occurrence of a set of causes; more specifically, unlike neurons diagrams and Petri Nets, causal Bayesian Networks use the conditional probability $P(B|A_1, \ldots, A_n)$ to capture the causal relationship between a set of causes $A_1, \ldots, A_n$ and an effect B. Bayesian networks are often represented by directed acyclic causal graphs (DAG), where each edge indicates existence of a conditional probability relationship between two variables.

Causal graphs: Causal graphs, where each edge denotes a causal relationship between a pair of variables, have been popularized by Pearl [133]. While they are similar to Bayesian networks, they are less expressive in that they do not capture joint causal relationships and cannot represent plural-causes. Nevertheless, they are one of the widely used models for representing causal knowledge, as they enable efficient algorithms for checking causal independence relationships among a given set of variables. However, since they are often constrained to DAGs, they cannot directly represent feedback (loops, self-causation).

(Fuzzy) cognitive maps: Cognitive maps [183] are similar to causal graphs, but are often enriched with more complex causal relationships, including positive (+), negative (−), neutral (0), causal effects [140]. Fuzzy cognitive maps (FCM) further introduce the concept of fuzziness factor to account for uncertainty and imprecision in causal relationships. One advantage of fuzzy cognitive maps is that, while being similar to Bayesian networks, they are not limited with probabilistic semantics: for example, fuzzy cognitive maps can be used to create multiple types of influence categories (e.g. positive influence (0,1], negative influence [-1,0), and neutral/no influence [0]). While FCMs allow negative causation, like causal graphs, they do not directly represent plural causes.

We note that, while there are multiple alternative causal frameworks, each providing different expressive powers (including the ability to represent temporal causal relationships in some cases), none of the existing frameworks have been designed with spatial causal relationships and we argue, in this paper, that this requires a new research agenda centered around space, time, and causal knowledge. In the next section, we start with the outlines of a spatio-temporal causal knowledge representation framework that overlays causal and spatial diffusion processes.

3.3 Causal Learning

Research on causal learning includes (1) leveraging of machine learning models to address fundamental causal inference tasks such as treatment effect estimation [105] and (2) causal discovery [166] as well as (3) leveraging of causality for improving machine learning tasks, such as interpretability [116] and fairness [143]. A consolidated summary of these causal learning tasks and the corresponding evaluation metrics are presented in Table 1 (more details are available in [30]). Below, we provide an overview of the major causal learning tasks:

- *Causal effect estimation* is the task of assessing the impact of an intervention (or treatment) on an outcome. Consider a medical study evaluating the effect of a vaccine on an epidemic causing virus. Patients are divided into two groups: one receiving the drug (treatment group) and one receiving a placebo (control group). Using metrics, such as Mean Absolute Error (MAE, Table 1), we can quantify how accurately we can estimate the vaccines effect on lowering transmission rate of the virus compared to the placebo. In this context, causality helps isolating the effect of the drug from other confounding factors, allowing us to attribute changes in transmission rate directly to the vaccine rather than other confounding variables. In the context of spatial data, however, the procedures and metrics in Table 1 need

		Causal Effect Estimation		Causal Structure Learning	Causal Interpretability and Fairness	
Metrics		**Standard Effect Metrics**	MAE, MSE, RMSE, PEHE, Policy Risk	SHD, SID, Frobenius Norm, Precision, Recall, F1, TPR, FPR, MSE, AUC, Precision-Recall Curve, FPR-TPR Curve, TVD, KL-Divergence, F-test	**Counterfactual Explanation**	Sparsity, Interpretability, Speed, Proximity, Diversity, Visual Linguistic
		Heterogenous Effect Metrics	$Uplift_{Coef}$, $Qini_{Coef}$		**Fairness**	FACE, FACT, Counterfactual Fairness, PC-Fairness, Ctf-DE, Ctf-IE, Ctf-SE
		Time Series Metrics	Standard and Heterogeneous Effect Metrics, F-Test, T-Test			
Procedures		**With Ground Truth**	Observational data with known effect; observational and experimental data pairs; sampling from observational data; sampling from synthetic data; sampling from RCTs	A transductive setting where we have the ground-truth causal graph and estimated graph	**Transductive**	Training on a regular dataset and testing on generated counterfactuals
		Without Ground Truth	Evaluation is possible if subset of the data is from RCTs		**Inductive**	Generating counterfactual explanations for an unseen instance

Table 1. An overview of causal learning tasks and metrics [30]

to be revised to account with causal effects across space and time (e.g. effect of school closures in one neighborhood to infection rates in others).

- While causal effect estimation focuses on a pair of treatment and intervention variables, *causal structure learning (or causal discovery)* is the process of identifying the causal structure underlying a complex system from observational data. Researchers for example might use historical public health data to determine if certain interventions, like masking, may reduce the rate of the epidemic growth. By applying causal discovery methods and metrics, such as the Structural Hamming Distance (SHD, Table 1), they can infer the directionality of the relationship (does masking lead to reduction in epidemics or do higher rates epidemics lead to increased mask usage?). While procedures and metrics in Table 1 focuses on more conventional causal discovery problems, in the context of spatio-causal settings, the discovered causal structures need to overlay on the underlying spatial system.
- *Causal interpretability* is the goal of providing explanations for predictions that consider the underlying causal relationships. Consider a public health decision support system which recommends vaccination of a sub-population. In this case, we may use a method like FACE (Fact-Counterfactual Fairness, Table 1) to explain this recommendation to the decision makers. The explanation could reveal that if this sub-population had a lesser baseline hospitalization rate, vaccination would not have been recommended, suggesting that the comorbidities suffered by this sub-population are a causal factor in the decision process. In this context, causality helps explore hypothetical scenarios and understand which factors are truly influencing the model's decisions and can improve fairness. Here, *causal fairness* refers to a set of techniques that aim to ensure that machine learning models make fair decisions. A vaccine recommendation system may potentially be biased against a certain population. Using techniques, such as counterfactual fairness, we can assess if an individual would receive the same healthcare resources if they belonged to a different population. Causality can assist in understanding and correcting for the biases present in the training data, aiming for consistent performance across different conditions. In spatio-causal settings, this can enable us to identify and correct biases in complex spatio-temporal systems.

In this paper, we argue that, in the context of spatio-temporal applications, research on causality has to start with the introduction of a *spatio-causal* framework as the current models for representing causal knowledge (briefly described next) are not expressive enough to effectively capture the various nuances and complexities introduced by causal inter-play between entities that are spatially distributed.

4 TOWARDS A SPATIO-TEMPORAL CAUSAL KNOWLEDGE REPRESENTATION FRAMEWORK

Integrating the physical and causal mechanisms impacting observed events and the interactions among system players is crucial when dealing with complicated real-world systems. It is imperative to have a thorough causal description framework. To handle uncertainty and bias, this framework needs to be able to capture socio-physical and spatio-temporal causal information while being domain-agnostic to support a variety of theories from the behavioral sciences to physics. The intricate spatio-temporal dynamics of complex systems are not well captured by existing approaches, such as causal graphs and structural equations [133, 134, 141]. Emergent behaviors, such the propagation of epidemics impacted by disease characteristics and therapies, are driven by the interactions and context-dependent causalities of interdependent entities found in these systems. In order to overcome these drawbacks, we provide a brand-new causal learning framework that is intended to faithfully replicate the complex physical processes found in real-world systems, enabling efficient data-driven forecasting and inference.

In particular, we define a complex system $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$ as a triple consisting of a spatial context, $\mathbb{S}$, a global context, $\mathbb{G}$, and a causal context, $\mathbb{C}$.

4.1 Spatial and Global Contexts

Definition 4.1 (Spatial Context). We define the *spatial context*, $\mathbb{S} = \{S, E, \Pi_s, \Pi_e, \Delta\}$, as a 5-tuple, where

- S is a set of spatial nodes,
- Π_s is the set of node properties – let $\pi_j^n \in \Pi_s$ be a node property; then $\pi_{i,j}^s = \pi_j^s(s_i)$ denotes the corresponding property for the spatial node $s_i \in S$,
- The set E of edges defined on the nodes in S describe a directed, edge labeled neighborhood graph, where Π_e is the set of edge properties. Let $\pi_j^e \in \Pi_e$ be an edge property and let $e_i = s_k \longrightarrow s_l$ be an edge in E. Given these, $\pi_{i,j}^e$ (also written as $\pi_j^e(e_i)$) denotes the property π_j^e of the edge e_i, and
- Δ is a set of distance functions – let $\delta_j \in \Delta$ be a distance function; then $\delta_{h,i,j}$ is the distance, under some distance concept, between any given pairs of nodes, $s_h, s_i \in S$ (note that we allow for multiple distance functions be defined over the same set of spatial nodes, each possibly governing the properties of the system based on different socio/physical processes). ◇

Intuitively, the spatial context of the complex system describes the spatial topology (described in terms of nodes, edges, and distance functions), along with the causally relevant node and edge properties.

Example 4.2. The following is a (simplified) example of a spatial context:

- $S = \{s1, s2, s3\}$, where $s1 = Mesa, s2 = Phoenix, s3 = Tempe$;
- $\Pi_s = \{Location, Altitude, Rain, Flood, Movement\}$;
- $E = \{e1, e2, e3, e4\}$, where $e1 = Phoenix \longrightarrow Tempe, e2 = Tempe \longrightarrow Phoenix, e3 = Tempe \longrightarrow Mesa, e4 = Mesa \longrightarrow Tempe$;
- $\Pi_e = \{Flow, Alt\_diff\}$; and
- $\Delta = \{\delta_{Euc}, \delta_{hop}\}$. †

In the above definition, a *property*, p, is a variable that takes values from a corresponding domain D_p. $A(p, x)$, then, indicates a *value assignment* of the form $p = x$ for the parameter p where $x \in D_p \cup \{\bot\}$. Here, $\bot$ indicates a null valued assignment (or lack of assignment) for that property.

Example 4.3. Let `annual_rain` be a property (or a variable) with domain $\mathbb{R}^+$. $A(annual\_rain, 2.35)$ is a possible value assignment. †

Given a spatial context, we further define a *variable* as a property whose value can potentially vary or be altered. In the rest of the paper, for simplicity, we will consider all properties as variables.

Spatial variables can be either spatially instantiated, meaning that they are specific to a particular node or edge, or not. Note that those variables that are not spatially instantiated can potentially refer to properties of any node or edge. The value that such a variable takes can be determined only after the variable is associated to a specific node or edge; i.e., after it gets spatially instantiated.

Definition 4.4 (Spatial Variables). Given a spatial context, $\mathbb{S} = \{S, E, \Pi_s, \Pi_e, \Delta\}$, the corresponding set, V_S, of *spatial variables* is defined as $V_S = V_S^{\otimes} \cup V_S^{\ominus}$, where $V_S^{\otimes}$ is the set of *spatially instantiated* node, edge, and distance variables

$$V_S^{\otimes} \quad = \quad V_{node}^{\otimes} \cup V_{edge}^{\otimes} \cup V_{dist}^{\otimes}, \quad \text{such that}$$

$$V_{node}^{\otimes} = \bigcup_{\pi_j^s \in \Pi_s \wedge s_i \in S} \pi_{i,j}^s, \qquad V_{edge}^{\otimes} = \bigcup_{\pi_j^e \in \Pi_e \wedge e_i \in E} \pi_{i,j}^e, \qquad V_{dist}^{\otimes} = \bigcup_{\delta_j \in \Delta \wedge s_k, s_l \in S} \delta_{j,k,l},$$

whereas $V_S^{\ominus} = V_{node}^{\ominus} \cup V_{edge}^{\ominus} \cup V_{dist}^{\ominus}$ is the set of node, edge, and distance variables that are *not spatially instantiated*:

$$V_{node}^{\ominus} = \bigcup_{\pi_j^s \in \Pi_s} \pi_{*,j}^s = \Pi_s, \qquad V_{edge}^{\ominus} = \bigcup_{\pi_j^e \in \Pi_e} \pi_{*,j}^e = \Pi_e, \qquad V_{dist}^{\ominus} = \bigcup_{\delta_j \in \Delta} \delta_{j,*,*} = \Delta.$$

◇

Example 4.5. The following is a set, V_S of spatial variables for the previous example:

$$\begin{aligned}
V_{node}^{\otimes} &= \{ \quad Mesa.Location, Mesa.Altitude, Mesa.Rain, Mesa.Flood, Mesa.Movement, \\
&\qquad Phoenix.Location, Phoenix.Altitude, Phoenix.Rain, Phoenix.Flood, Phoenix.Movement, \\
&\qquad Tempe.Location, Tempe.Altitude, Tempe.Rain, Tempe.Flood, Tempe.Movement \quad \}, \\
V_{edge}^{\otimes} &= \{ \quad e_1.Flow, e_2.Flow, e_3.Flow, e_1.Alt_diff, e_2.Alt_diff, e_3.Alt_diff \quad \}, \\
V_{dist}^{\otimes} &= \{ \quad \delta_{Euc}(Mesa, Phoenix), \delta_{Euc}(Phoenix, Mesa), \delta_{Euc}(Tempe, Phoenix), \\
&\qquad \delta_{Euc}(Phoenix, Tempe), \delta_{Euc}(Mesa, Tempe), \delta_{Euc}(Tempe, Mesa) \quad \}, \\
V_{node}^{\ominus} &= \{ \quad Location, Altitude, Rain, Flood \quad \}, \\
V_{edge}^{\ominus} &= \{ \quad Flow, Alt_diff \quad \}, \\
V_{dist}^{\ominus} &= \{ \quad \delta_{Euc}, \delta_{hop} \quad \}.
\end{aligned}$$

†

A *global context*, $\mathbb{G}$, is defined as a set, V_G, of global properties and variables not associated to any spatial nodes or properties of a given spatial context $\mathbb{S}$.

Example 4.6. For our running example, let the global context be defined based on the set

$$V_G = \{Season, News, Regional\_damage\}.$$

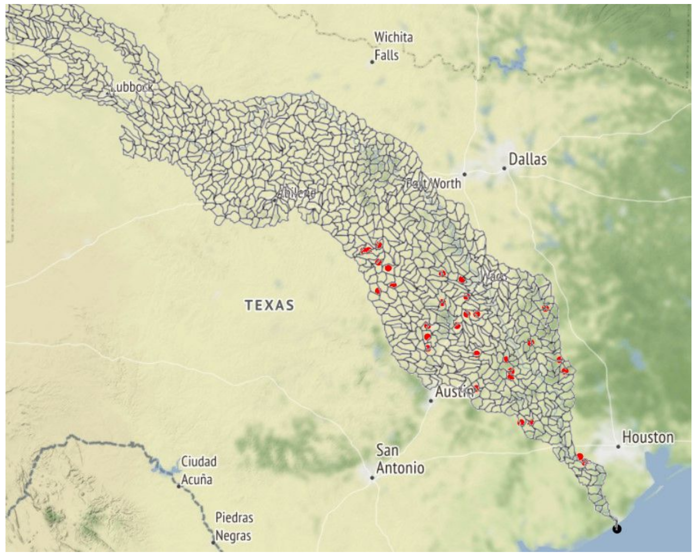


Fig. 7. Causally relevant nodes, in the context of a stream-flow estimation task, selected by the STREAMS [164] algorithm for a target (shown in black). The highlighted regions indicate the watersheds upstream from the target, according to the HAWQS model [1], employed by the EPA and USDA. In this example, the amount of water at a particular watershed has a *diffusing* impact on downstream watersheds, as a function on the watershed network / river connectivity as well as hydrological properties of the connecting watersheds

†

4.2 Causal Context

The causal context of the complex system describes the causal constraints that apply on the variables of its spatial and global contexts. These causal constraints themselves can have properties, or *causal attributes* that are distinct from the spatial and global properties and quantify various aspects of the corresponding causal relationships. Examples include causal strength, σ (indicating the severity of an effect), causal lag λ (indicating the time difference between a cause and the resulting effect), and causal likelihood, ω (indicating the probability of observing an effect given the cause).

Let the set, Π_c, be the set of of *causal attributes*.

As before, $A(\Pi_c, X)$ denotes an assignment for those causal properties from a value within their respective domains or $\perp$ (signifying the lack of an assignment for that particular causal property) – X denotes the corresponding assignment values The causal context then includes a set of causal statements of different types: (a) statements with node centric effects; (b) statements with edge centric effects, (c) statements with global effects, and (d) statements with effects on causal attributes, such as strength and lag, associated with other causal statements. We formally define these statements below.

4.2.1 *Causal Statements with Node Centric Effect.*

Causal statements with node centric effect are causal statements which describe cause-effect relationships where the effect is on a node property. We define two types of causal statements with node-centric effect: **spatially focused** and **spatially diffused**. Intuitively, a focused causal statement with node-centric effect indicates that the effect is limited to the property of a single node. A diffused spatial causal statement on the other hand, indicates that the causal impact is not limited to a single node, but *may spread to other nodes on the underlying spatial network*, following the outgoing spatial edges (see Figure 7 for an example).

Definition 4.7 (Causal Statement with a Node-Centric Effect). Let $v_i \in V_S$ and let $v_j \in V^{\otimes}_{node} \cup V^{\ominus}_{node}$. A *focused causal statement with node-centric effect* from v_i to v_j is denoted as

$$e = v_i \xrightarrow[sf]{A(\Pi_c,X)} v_j,$$

whereas a *diffused causal statement with node-centric effect* is denoted as

$$e = v_i \xrightarrow[sd]{A(\Pi_c,X)} v_j.$$

Here v_i is a spatial (node or edge) or global variable that causes an effect on the node variable v_j. The cause-effect relationship is regulated by the specified assignment $A(\Pi_c, X)$ for the relevant causal attributes.

◇

For simplicity of discussion, in the rest of this section, we will ignore the concept of causal likelihood – therefore each edge will be associated with a single lag and a single strength. In reality though, each causal statement would be associated with a set of lag/strength pairs, each with a different likelihood.

Example 4.8. The following is a spatially focused causal statement with node-centric effect:

$$c_1: \quad \texttt{rain} \xrightarrow[sf]{\lambda = 2, \sigma = \perp} \texttt{flood}$$

In this example, the causal effect described by this causal statement is spatially focused in that the rain causes flood in the spatial location where it is recorded – the σ of the causal relationship has not been specified. †

Note that, as we see in the next example, if only the source or the destination spatial location is specified, it is assumed to denote both the source and destination of the causal statement.

Example 4.9. The following is a spatially focused causal statement with node-centric effect:

$$c_2: \quad \texttt{rain} \xrightarrow[sf]{\lambda = \perp, \sigma = \perp} \texttt{Mesa.flood}$$

In this example, the causal effect described by this causal statement is spatially focused in that the rain in Mesa causes flood in the same location and the λ and σ of the causal relationship have not been specified. †

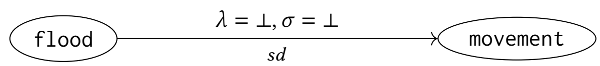
Example 4.10. The following is a spatially diffused causal statement with node-centric effect:

$$c_3: \quad \texttt{Mesa.rain} \xrightarrow[sd]{\lambda = \perp, \sigma = \perp} \texttt{Tempe.flood}$$

In this example, the causal effect is spatially instantiated in that the rain in "Mesa" would cause flood in "Tempe" with an unspecified lag. However, the specified causal effect is spatially transitive, meaning that any spatial node downstream from "Tempe" can potentially be causally effected from the rain in "Mesa". †

Example 4.11. Consider a scenario where occurrence of a flood causes population movement in the spatial location where it is recorded as well as nearby locations. We would represent this with the following spatially diffused causal statement:

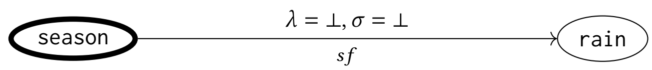
c_4:



†

Example 4.12. The following is a spatially focused causal statement with node-centric effect:

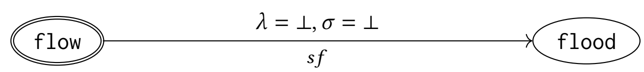
c_5:



In this example, season (which is a global variable) impacts the amount of rain with an unspecified lag and strength. Note that, in this example, we use a thick border to visually indicate a global property/variable. †

Example 4.13. The following is a spatially focused causal statement with node-centric effect:

c_6:



In this example, waterflow on a spatial edge causes flood on the destination location, with an unspecified lag and strength. Here, we use an ellipse with a double-border to visually indicate an edge variable. †

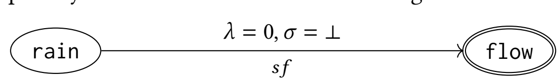
4.2.2 *Causal Statements with Edge Centric Effect.*

Causal statements with edge centric effect are causal statements which describe cause-effect relationships where the effect is on an edge property. We again define two types of causal statements with edge-centric effect: spatially focused and spatially diffused: a focused causal statement with edge-centric effect indicates that the effect is limited to the property of a single edge. A diffused spatial causal statement on the other hand, indicates that the causal impact is not limited to a single edge, but may spread to other downstream edges on the underlying spatial network.

Definition 4.14 (Causal Statement with Edge-Centric Effect). A *causal statement with edge-centric effect* is defined similarly to a node-centric causal statements, except that $v_j \in V^{\otimes}_{edge} \cup V^{\ominus}_{edge}$. ◇

Example 4.15. The following is a spatially focused causal statement with edge-centric effect:

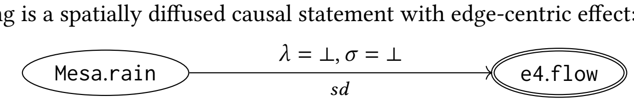
c_7:



In this example, the causal effect described by this causal statement is spatially focused in that the rain causes waterflow on the edges that are outgoing from the node where it has been recorded with 0 lag; the strength of the causal relationship has not been specified. †

Example 4.16. The following is a spatially diffused causal statement with edge-centric effect:

c_8:



In this example, the rain recorded in Mesa causes waterflow on the edge e_4 (which is from Mesa to Tempe) as well as all downstream edges; the lag and the strength of the causal relationship have not been specified. †

Example 4.17. The following is a spatially diffused causal statement with edge-centric effect:

$$c_9: \quad \texttt{e1.flow} \xrightarrow[sd]{\lambda = \perp, \sigma = \perp} \texttt{e3.flow}$$

In this example, the waterflow on edge e1 (which is from Phoenix to Tempe) impacts waterflow on edge e3 (which is from Tempe to Mesa) as well as all downstream edges. The λ and σ of the causal relationship have not been specified. †

Example 4.18. The following is a spatially focused causal statement with edge-centric effect:

$$c_{10}: \quad \texttt{alt\_diff} \xrightarrow[sf]{\lambda = \perp, \sigma = \perp} \texttt{flow}$$

This causal statement specifies that the altitude difference between two locations (recorded in the Alt_diff property on the edge from one location to the other) causally impacts the water flow from the source location to the destination location. †

4.2.3 *Causal Statements with Global Effect.*

Causal statements with global effect are causal statements which describe cause-effect relationships where the effect is on a global property. Note that since these statements apply globally they are not delineated as spatially focused or spatially diffused.

Definition 4.19 (Causal Statement with Global Effect). These are defined similarly to node- and edge-centric causal statements, except that $v_j \in V_G$ and causal statements are not delineated as being spatially focused or diffused. ◇

Example 4.20. The following is a causal statement with global effect:

$$c_{11}: \quad \texttt{flood} \xrightarrow{\lambda = \perp, \sigma = \perp} \texttt{regional\_damage}$$

This causal statement specifies that any flood at any spatial location contributes to the overall regional damage. †

4.2.4 *Causal Statements with Effect on Causal Attributes of Other Causal Statements.*

The causal statements we have considered so far, all, describe effects on spatial or global properties in the complex system. These type of statements alone, however, fail to capture causality that is context dependent. For instance, consider a scenario, where the lag with which the flood propagates over an edge is determined by the soil properties. Here, the soil property does not directly act on the flood at the destination node, but acts on the lag attribute of the causal statement that describes the flood process. We, therefore, also need causal statements whose effects are causal attributes of other causal statements in the causal context.

Definition 4.21 (Causal Attribute Variables). Let C be a set of causal statements. We define the corresponding set, V_C of *Causal Attribute Variables* as follows:

$$V_C = \{\pi^c_j(c_i) \mid \pi^c_j \in \Pi_c \wedge c_i \in C\},$$

where $\pi^c_j(c_i)$ denotes the causal attribute π^c_j of causal statement c_i. ◇

Causal statements with effect on causal attributes of other causal statements then act on the causal attribute variables of causal statements that have been already specified.

Definition 4.22 (Causal Statement with Effect on a Causal Attribute). A *causal statement with effect on a causal attribute* is defined similarly to node- and edge-centric causal statements, except that

$$\begin{aligned} v_i &\in V^{\ominus}_{node} \cup V^{\ominus}_{edge} \cup V^{\ominus}_{dist} \cup V_G \cup V^{spec} \quad \text{where} \\ V^{spec} &= \{n\_spec : v_h \mid n\_spec \in NSpec \wedge v_h \in V^{\ominus}_{node}\} \cup \\ &\quad\ \{e\_spec : v_h \mid e\_spec \in ESpec \wedge v_h \in V^{\ominus}_{edge}\} \cup \\ &\quad\ \{n\_spec_1 : n\_spec_2 : v_h \mid n\_spec_1 \neq n\_spec_2 \in NSpec \wedge v_h \in V^{\ominus}_{dist}\}. \end{aligned}$$

and $v_j \in V_C$. Above, $NSpec$ and $ESpec$ are node and edge specifiers that help resolve ambiguities in the specification, when there are multiple nodes or edges in the causal statement. More specifically, the spatial node specifiers in the set

$$NSpec = \{c, e, c : s, e : s, c : d, e : d\}$$

help distinguish among nodes referred to in the *cause* (c) or *effect* (e) clause of the target causal statement; when the causal statement include edges, however, the node specifiers help distinguish among *source node of a cause or effect* ($c : s$ or $e : s$) and *destination node of a cause or effect* ($c : d$ or $e : d$). The spatial edge specifiers in

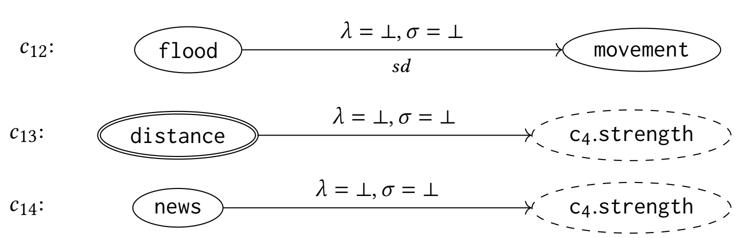
$$ESpec = \{c, e\}$$

on the other hand help distinguish among edges referred in the *cause* (c) or *effect* (e) clause of the target causal statement. ◇

As with the causal statements with global effect, these causal statements are not delineated as being spatially focused or diffused.

Example 4.23. Consider a scenario where occurrence of a flood at a particular location causes movement of people in the spatial location where it is recorded and nearby nodes. The population movement in the event of a hurricane is (not caused, but) influenced by the distance between the locations and the nature of news carried by the media.

Conventional causal graphs cannot adequately capture scenarios where floods cause population movements, influenced by factors like the distance between locations and media influence. They fall short in representing spatial diffusion, the network's topology effects, and the nuanced influences on causal relationships, such as how news and distances between nodes modify the impact of a flood on population movement. In the proposed formulation, however, this would be represented as



†

As can be seen above, the framework allows a richer expression of the scenario which cannot be formulated using the existing causal representations that are uniformed by the spatial processes and cannot accurately distinguish causes from influencers. A particular advantage of the proposed causal statements with effects on a causal attributes is the definition of inhibitors, where the effect of a particular global or spatial property may be the inhibition of a particular cause-effect relationship by adjusting its strength (or likelihood) to 0. For example, in the above example, a particularly close distance between a pair of nodes may have an encouraging whereas a particularly large distance between a pair of nodes may have an inhibiting effect on the population movement.

Below, we include several additional examples to highlight this important class of causal statements.

Example 4.24. The following is a causal statement with effect on a causal attribute:

$$c_{15}: \quad \text{season} \xrightarrow{\lambda = \perp, \sigma = \perp} c_3.\lambda$$

This causal statement specifies that seasonal conditions may impact the lag of the causal statement c_3; i.e., lag with which rain in "Mesa" may cause flood in "Tempe" (and transitively on reachable locations from "Tempe"). †

Example 4.25. The following is a causal statement with effect on a causal attribute:

$$c_{16}: \quad \delta_{\text{Euc}} \xrightarrow{\lambda = \perp, \sigma = \perp} c_1.\lambda$$

This causal statement specifies that Euclidean distance between the two nodes involved in the causal statement c_1 may impact the lag of that causal statement; i.e., the lag would depend on the Euclidean distance between the spatial location at which rain is recorded and the location where the flood occurs). †

In this figure, we use a dotted borderline to visually mark distance variables.

Example 4.26. The following is a causal statement with effect on a causal attribute:

$$c_{17}: \quad \text{flow} \xrightarrow{\lambda = \perp, \sigma = \perp} c_6.\sigma$$

This causal statement specifies that the amount of flow involved in the causal statement c_6 also impacts the corresponding causal strength. †

Example 4.27. The following is a causal statement with effect on a causal attribute:

$$c_{18}: \quad \text{rain} \xrightarrow{\lambda = \perp, \sigma = \perp} c_7.\sigma$$

This causal statement specifies that the amount of rain recorded at a location also impacts the causal strength of the causal statement c_7 (which states that rain at a given location and outgoing flow are causally related. †

Example 4.28. The following is a causal statement with effect on a causal attribute:

$$c_{19}: \quad \text{c:s:rain} \xrightarrow{\lambda = \perp, \sigma = \perp} c_{10}.\sigma$$

This causal statement specifies that the amount of rain recorded at the source location of the spatial edge referred to as the cause of c_{10} also impacts the causal strength of the causal statement. †

Example 4.29. The following is a causal statement with effect on a causal attribute:

$$c_{20}: \quad \text{e:d:flood} \xrightarrow{\lambda = \perp, \sigma = \perp} c_{10}.\sigma$$

This causal statement specifies that the amount of flood already recorded at the destination location of the spatial edge referred to as the effect of c_{10} also impacts the causal strength of the causal statement. †

Definition 4.30 (Causal Context). Given a spatial context, $\mathbb{S} = \{S, E, \Pi_s, \Pi_e, \Delta\}$, a global context $\mathbb{G}(= V_G)$, and a set Π_c of causal attributes, a *causal context*, $\mathbb{C}$, is defined as a set of causal statements with respect to the given spatial/global context. ◇

Example 4.31. The following is a causal context in our running example.

$$\mathbb{C} = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}, c_{11}, c_{12}, c_{13}, c_{14}, c_{15}, c_{16}, c_{17}, c_{18}, c_{19}, c_{20}\}.$$

†

Note that, above, we intentionally use a simplified formulation of a causal context, where each causal statement describes a socio/physical process that independently acts on its effect (but possibly regulated by other causal statements with effects on its causal attributes). In reality, though, certain effects may appear *only when multiple spatial causes co-exist* or other effects may appear *if and only if one or the other potential causes exist*, but not if both exist simultaneously. Therefore, unlike basic causal graphs which have binary edges, the causal context will also need to account for conjunctive and disjunctive statements (such as $rain \wedge surface\_conditions \longrightarrow flood$ vs. $rain \vee surge \longrightarrow flood$). In this paper, we ignore these more complex spatio-causal relationships and leave them for further work that will study them in greater detail. Note that the proposed formalism, nevertheless, is richer in its expressive power than conventional causal graphs as it overlays a causal graph over a dynamic spatio-temporal network and its statements operate not only on variates in the system, but also on other causal statements. This richer language will, therefore, require investigation and re-definition of key causal concepts (such as mediators and confounders, foreward and backwards paths, Markov blankets, and d-separation), as well as decounfounding algorithms critical for causal discovery, data imputation, and forecasting (see Section 3).

4.3 Complex Spatial System and spatio-temporal Observations

A *complex spatial system*, $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$ consists of a

- a spatial context, $\mathbb{S}$,
- a global context, $\mathbb{G}$, and
- a causal context, $\mathbb{C}$.

Given a complex spatial system, $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$, an *observation instance*, O, of $\mathfrak{C}$ is a multi variate time series, $\mathbb{T}(t_{start}, t_{end})$, where for each observable property, $p \in P_O = P_S^{\otimes} \cup P_G$, $T_i \in \mathbb{T}(t_{start}, t_{end})$ is a time series starting at t_{start} and ending at t_{end} such that, for all $t \in [t_{start}, t_{end}]$, we have $p[t] \in D_p \cup \{\perp\}$, where $\perp$ indicates the lack of an observation for the property, p, at the given time t. Note that some of the properties may be subject to additional set, $\mathbb{A}$, of assumptions/constraints. For example, let $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$ be a complex spatial system and let $p \in P_O = P_S^{\otimes} \cup P_G$ be an observable property. The *temporal immutability* constraint, I, on p would impose that, if observed, the value of $p[t]$ is constant across all values of t.

4.3.1 *Causal Consistency of a Complex Spatial System with a Set of Observations.*

We say that a complex spatial systems and observations are consistent if there are no causal fallacies or violations of the spatio-temporal assumptions/constraints. Let us be given

- a complex spatial system, $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$,
- subject to a set, $\mathbb{A}$, of assumptions/constraints, and
- a set, $\mathbb{O}$, of observation instances.

We say that the triple $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$ is consistent (i.e., $(\mathfrak{C}, \mathbb{A}, \mathbb{O}) = true$) iff the following conditions hold:

- The causal context, $\mathbb{C}$, does not include any causal fallacies[1] (such as causal cycles).
- The triple, $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$, does not include any temporal inconsistencies: in other words, for all p_i in the spatial or global context and for all t with a non-null observation $p_i[t] = x$,

$$\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \setminus \{p_i[t]\} \rangle \nRightarrow p_i[t] = y,$$

  where $y \neq x$. In cases where the triple $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \setminus \{p_i[t]\} \rangle$ could imply multiple alternative values for $p_i[t]$, we would seek to maximize the likelihood of $p_i[t]$ being equal to x (ideally, $prob(p_i[t] = x) = 1$).
- The triple, $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$, does not include any inconsistencies violating assumptions in the set $\mathbb{A}$: in other words, for all p_i in the spatial or global context and for all t without an observation or with a null observation ($p_i[t] = \perp$),

$$\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle \nRightarrow p_i[t] = y,$$

  where $p_i[t] = y$ and observation instances $\mathbb{O}$, together, would violate the given set of assumptions/constraints. In cases where the triple $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$ could imply multiple alternative values for $p_i[t]$, it should be that for any assumption/constraint on the property p_i,

$$\exists_{x \in D_p} \prod_t prob(p_i[t] = x) > 0.$$

4.3.2 *Causal Compatibility of a Causal Statement with a Complex Spatial System under a Given set of Observations.*

Given this, we then define causal compatibility of a causal statement with a complex spatial system in terms of the causal compatibility of the extended causal context under a given set of observations and spatio-temporal assumptions/constraints, Let us be given a complex spatial system, $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$, subject to a set, $\mathbb{A}$, of assumptions/constraints, and a set, $\mathbb{O}$, of observation instances. Let us assume that the triple $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$ is consistent. Let us further be given a causal statement

$$e = v_i \xrightarrow[^{sf} \text{ or } ^{sd} \text{ or NA}]{A(\Pi^c, X)} v_j$$

and let $\widehat{\mathfrak{C}} = \{\mathbb{S}, \mathbb{G}, \widehat{\mathbb{C}}\}$ be the extended complex spatial system, where

$$\widehat{\mathbb{C}} = \mathbb{C} \cup \{e\}.$$

We say that e is compatible with the triple, $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$ iff the triple $\langle \widehat{\mathfrak{C}}, \mathbb{A}, \mathbb{O} \rangle$ is causally consistent:

$$compatible(e, \langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle) = consistent(\widehat{\mathfrak{C}}, \mathbb{A}, \mathbb{O}).$$

4.4 Spatio-Causal Inference and Discovery

Given a complex spatial system $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$ and a set of observation instances, $\mathbb{O}$, with possibly null values, or values that represent ranges or probability distributions, along with additional assumptions, $\mathbb{A}$ (such as temporal immutability

[1]Note that checking for causal fallacies requires a framework in which causal contexts can be interpreted. Examples include Bayesian [136] and Petri Net based [144] definitions of causal consistency.

constraints on a subset of the variables), we can then define various socio/spatio-temporal imputation and forecasting problems as well as spatio-temporal causal discovery tasks.

4.4.1 Spatio-Causal Inference Tasks.

Problem 4.32 (spatio-temporal Imputation Problem). Let us be given a complex spatial system, $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$, subject to a set, $\mathbb{A}$, of assumptions/constraints, a set, $\mathbb{O}$, of observation instances, and a definition of causal consistency. Let $O_i \in \mathbb{O}$ be an observation instance and for some $t \in [t_{start,i}, t_{end,i}]$ let $p[t] = \bot$ for some observable property p. The corresponding imputation problem aims to recover the value of $p[t]$; i.e., the task is to identify and return the value y, s.t.

$$\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle \Rightarrow p[t] = y.$$

In other words, the triple $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$ would imply that the value of the property p at time t is equal to y. In cases where the triple $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$ could imply multiple alternative values for $p[t]$, we would aim to return either a probability distribution for the values of y, or the most likely value $\tilde{y}$ given the underlying probability distribution. ‡

Problem 4.33 (spatio-temporal Forecasting Problem). Let us be given a complex spatial system, $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$, subject to a set, $\mathbb{A}$, of assumptions/constraints, a set, $\mathbb{O}$, of observation instances, and a definition of causal consistency. Let $O_i \in \mathbb{O}$ be an observation instance. The forecasting problem aims to predict the value of $p[t]$, for some observable property p, for some target $t > t_{end,i}$; i.e., the task is to identify and return the value y, s.t.

$$\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle \Rightarrow p[t] = y.$$

As in the case of the imputation problem, in cases where the triple $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$ could imply multiple alternative values for $p[t]$, we would aim to return either a probability distribution for the values of y or the most likely value $\tilde{y}$ given the underlying probability distribution. ‡

4.4.2 Spatio-Causal Discovery Task.

In order to formalize the spatio-temporal discovery task, we first need to introduce a parametrized causal statement that will serve as the blueprint for the discovery tasks.

Definition 4.34 (Parametrized Causal Statement). Let us be given a complex spatial system, $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$, subject to a set, $\mathbb{A}$, of assumptions/constraints, and a set, $\mathbb{O}$, of observation instances. A *parametrized causal statement*, $statement(\Theta)$, is of the form

$$statement(\Theta) = q_{cause} \xrightarrow[sf \text{ or } sd \text{ or NA}]{q_{c\_prop}} q_{effect},$$

and uses the symbol "?" to indicate free parameters. More specifically, $statement(\Theta)$, is

- either a parametrized causal statement with node or edge centric effect (with free node or edge parameters in its cause and/or effect clause); i.e., both q_{cause} and q_{effect} are in

$$\left\{?\right\} \cup \left(\bigcup_{(j=?)\vee(\pi_j^g \in \Pi_G)} \pi_j^g \right) \cup \left(\bigcup_{\substack{((j=?)\vee(\delta_j \in \Delta))\wedge \\ (((\langle k,l\rangle \in \{\langle *,*\rangle,\langle ?,?\rangle\})\vee(s_k,s_l \in S))}} \delta_{j,k,l} \right) \cup \left(\bigcup_{\substack{((j=?)\vee(\pi_j^s \in \Pi_S))\wedge \\ ((i\in\{*,?\})\vee(s_i \in S))}} \pi_{i,j}^s \right) \cup \left(\bigcup_{\substack{((j=?)\vee(\pi_j^e \in \Pi_e))\wedge \\ ((i\in\{*,?\})\vee(e_i \in E))}} \pi_{i,j}^e \right).$$

- or a parametrized causal statement with effect on a causal attribute (with free node or edge parameters in its cause clause); i.e., q_{cause} is in

$$\left\{?\right\} \cup \left(\bigcup_{(j=?)\vee(\pi_j^g \in \Pi_G)} \pi_j^g \right) \cup$$
$$\left(\bigcup_{(j=?)\vee(\pi_j^s \in \Pi_S)} \pi_{*,j}^s \right) \cup \left(\bigcup_{(j=?)\vee(\pi_j^e \in \Pi_e)} \pi_{*,j}^e \right) \cup \left(\bigcup_{((j=?)\vee(\delta_j \in \Delta))} \delta_{j,*,*} \right) \cup$$
$$\left(\bigcup_{\substack{(x \in NSpec)\wedge \\ ((j=?)\vee(\pi_j^s \in \Pi_S))}} x : \pi_{*,j}^s \right) \cup \left(\bigcup_{\substack{(x \in ESpec)\wedge \\ ((j=?)\vee(\pi_j^e \in \Pi_e))}} x : \pi_{*,j}^e \right) \cup \left(\bigcup_{\substack{(x,y \in NSpec)\wedge \\ ((j=?)\vee(\delta_j \in \Delta))}} x : y : \delta_{j,*,*} \right)$$

while q_{effect} is in

$$\left(\bigcup_{(\pi^c \in \Pi_c)\wedge(c \in \mathbb{C})} \pi^c(c) \right).$$

Above, the symbol "?" indicates a free parameter without a bound value and q_{c_prop} is a set of free/bound value assignments for the corresponding causal properties, such that for $p \in \Pi_c$, we have $p = x$ where $x \in D_p \cup \{\perp, ?\}$. If the statement has no free variables, i.e., $\Theta = \emptyset$, then the statement is referred to as a *bound statement*.

Note that the specification for the parametrized causal statement with effect on a causal attribute is more complex as, per Definition 4.22, the cause clause of such a statement may contain node or edge specifiers, which may themselves be parametrized.

◇

Example 4.35. The following is a parametrized causal statement:

$$statement_1 : \quad ? \xrightarrow{\lambda = \perp, \sigma = \perp} c_3.\lambda$$

This statement states that a parametrized global parameter may impact the lag of the causal statement c_3.

†

Example 4.36. The following is a spatially focused causal statement with node-centric effect:

$$statement_2 : \quad \texttt{flow} \xrightarrow[sf]{\lambda = ?, \sigma = \perp} ?$$

In this example, waterflow on a graph edge causes a spatially focused impact on some parametrized property of the destination location, with a parametrized lag and an unspecified strength. †

Example 4.37. The following is a spatially transitive causal statement with node-centric effect:

$$statement_3 : \quad \texttt{?.rain} \xrightarrow[sd]{\lambda = \perp, \sigma = \perp} \texttt{Tempe.flood}$$

In this example, rain at a parametrized node causes flood in Tempe (and potentially other cities reachable from Tempe).

†

Problem 4.38 (Spatio-Causal Learning Task or Spatio-Causal Query). Let us be given a complex spatial system, $\mathfrak{C} = \{\mathbb{S}, \mathbb{G}, \mathbb{C}\}$, subject to a set, $\mathbb{A}$, of assumptions/constraints, and a set, $\mathbb{O}$, of observation instances. Let us assume that the triple $\langle \mathfrak{C}, \mathbb{A}, \mathbb{O} \rangle$ is consistent.

A *spatio-causal query*, $q_{spatio\_causal}(\Theta) = statement(\Theta)$, seeks a set of assignments, ϕ, to the free parameters of a given parametric causal statement that collectively render the corresponding bound statement, $statement(\Theta = \phi)$, compatible with the given complex spatial system (including the underlying, partial, causal context $\mathbb{C}$), temporal immutability constraints, and the observation instances. In cases where there are multiple sets of assignments to the free parameters of a given parametric causal statement that collectively render the corresponding bound statement compatible with the given complex spatial system, we would aim to return either a probability distribution for the assignments, or the most likely assignment, $\tilde{\phi}$, given the underlying probability distribution.

‡

Given the key concepts introduced above, in the next section, we will start discussing approaches to various spatio-temporal causal learning (inference and/or discovery) problems.

5 SPATIO-TEMPORAL CAUSAL LEARNING

Humans undergo a complex process known as causal learning to comprehend the effects within a system, a concept similarly applied in machine learning to enhance model performance through causality, with applications from recommender systems [162] to autonomous driving [157]. Causal research in machine learning is divided into causal discovery [171] and causal inference [135]:

- To enhance the understanding of causal concepts, it is essential for the models to discover which variables follow a causal relationship amongst themselves. The task of discovering causal relationships from any form of data is known as causal discovery.
- Once we identify the causal relationships, the next step is to understand how these relationships are affected and estimate the effect of different interventions. The task of estimating the effects is known as causal inference.

We emphasize the need for spatio-temporal causal models in diverse fields like sustainability, public health, climate science, and social sciences, to understand causal linkages in space and time. These models are crucial for analyzing regionally dispersed variables in time series data, recognizing causative variables, and determining causation directions, considering spatial and temporal dependencies.

5.1 An Example: Water Runoff Rate Prediction

The comprehension of the fluctuation in causal connections throughout various times and geographical locations, particularly in relation to external occurrences, presents a notable obstacle. Consider, for instance, the endeavor of forecasting runoff rates to improve water sustainability, as previously said. The process of runoff rate prediction entails the estimation of the quantity of surface water runoff originating from a designated region over a given period, usually subsequent to precipitation events. The management of water resources in regions experiencing scarcity or where demand exceeds supply is of utmost importance. The ability to make precise forecasts empowers water resource managers to efficiently allocate and distribute resources, adopt flood control strategies, and mitigate the potential for resource scarcity.

The concept of spatio-temporal causality is of utmost importance in the prediction of runoff rates in different geographical areas. It involves identifying the causative elements that contribute to variations in runoff rates throughout different time periods and spatial locations. By doing spatio-temporal causal analysis, it is possible to determine the direction and magnitude of causal connections between runoff rates, identify significant causal elements, and comprehend their interaction within certain geographical regions. This observation enables the formulation of targeted

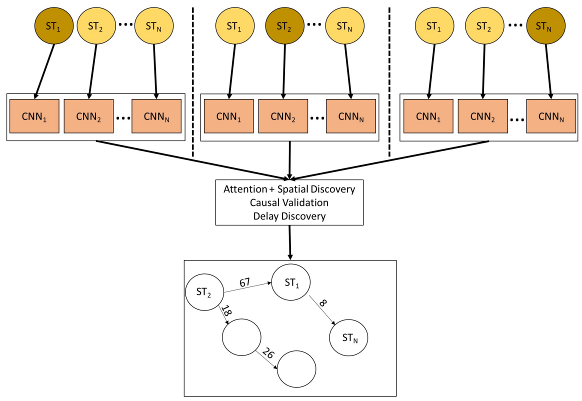


Fig. 8. Overview of STCD architecture, showing the process from inputting time-series and geographical data (ST-nodes) through N CNNs, to creating a causal network graph with edges indicating timestep delays. A directed edge from ST_1 to ST_N signifies a causal relationship from ST_1 to ST_N.

actions aimed at reducing runoff rates and improving water sustainability. For example, in the event that it is determined that alterations at location A have a substantial impact on the rates of runoff at target location C, it becomes possible to strategically implement interventions that aim to enhance infiltration and mitigate surface runoff.

Let us be given N locations and let $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,T})$ represents the readings for the i^{th} location over the past T units of time. Let D represent the spatial distance matrix, where $D_{i,j}$ represents the distance between location i and j, where the sign of $D_{i,j}$ is positive if location i is geographically higher than location j and negative otherwise. The objective is to reconstruct the input based on the previous readings of each location and to infer a causal graph G of size $N \times N$, taking into account the underlying spatial diffusion process. Inferring the underlying causal graph can aid in identifying the crucial locations that affect the runoff at a target location and lead to better generalizability across time and better interpretability.

5.2 Data Structures, Models, and Algorithms for Causal Discovery

spatio-temporal causal discovery aims to identify causal relationships among geographically dispersed variables over time. By analyzing spatial and temporal data relationships, these models reveal causal mechanisms influencing variables' behaviors. For instance, they enable the identification of key geographic factors affecting runoff rates at specific locations through the analysis of topography and geographical data across different regions and timeframes. This knowledge assists in refining predictive models and formulating strategies for water sustainability.

STCD employs an architecture of n-CNNs, processing each sensor's data with attention mechanisms to discern temporally and spatially significant features. Unlike temporal methods that rely solely on event sequences, STCD integrates spatial dimensions, essential for understanding causality in settings where geographical relationships affect outcomes. This approach overcomes limitations of previous methods by considering the directional and magnitude aspects of spatial relationships, essential in contexts like hydrology where the spatial arrangement of causes relative to their effects significantly influences causal inference.

Prior research has established foundational models for causal discovery in static and temporal settings. However, these models often fall short in spatio-temporal contexts, where the interplay between space and time complicates causal

Model		No. of features	RAE Year 2006	RAE Year 2007	RAE Year 2008	SMAPE Year 2006	SMAPE Year 2007	SMAPE Year 2008	MSE Year 2006	MSE Year 2007	MSE Year 2008
TCDF	**LSTM**	1,356	0.99	0.97	1.07	1.69	1.37	1.49	0.15	0.09	0.21
All Features		3,128	1.10	1.03	1.29	1.72	1.42	1.58	0.18	0.11	0.24
Random Features		1,248	1.02	1.07	1.13	1.76	1.48	1.54	0.16	0.13	0.26
Remaining features		1,880	1.04	1.05	1.10	1.74	1.46	1.52	0.17	0.11	0.29
STCD		1,248	**0.95**	**0.91**	**1.02**	**1.66**	**1.31**	**1.43**	**0.12**	**0.06**	**0.14**
TCDF	**CNN**	1,356	0.97	0.89	1.13	1.63	1.31	1.61	0.13	0.09	0.16
All Features		3,128	1.02	0.94	1.20	1.65	1.34	1.65	0.16	0.12	0.21
Random Features		1,248	1.05	0.96	1.24	1.69	1.37	1.71	0.14	0.14	0.27
Remaining features		1,880	1.04	0.94	1.21	1.66	1.35	1.63	0.14	0.12	0.25
STCD		1,248	**0.92**	**0.84**	**1.05**	**1.59**	**1.28**	**1.56**	**0.10**	**0.05**	**0.12**
TCDF	**SVR**	1,356	1.28	1.10	1.78	1.87	1.68	1.87	0.24	0.19	0.31
All Features		3,128	1.33	1.12	1.83	1.89	1.72	1.92	0.29	0.23	0.34
Random Features		1,248	1.38	1.17	1.86	1.87	1.77	1.95	0.26	0.20	0.39
Remaining features		1,880	1.39	1.15	1.84	1.88	1.74	1.93	0.27	0.22	0.36
STCD		1,248	**1.20**	**1.07**	**1.71**	**1.84**	**1.60**	**1.81**	**0.18**	**0.14**	**0.27**

Table 2. Prediction performances obtained using different combinations of the features across three different years. The results in bold represent the best results (details of the experimental setup and the discussions are available at [166])

Model	No. of Features	Year 2000 RMSE	Year 2000 MASE	Year 2000 SMAPE	Year 2004 RMSE	Year 2004 MASE	Year 2004 SMAPE	Year 2008 RMSE	Year 2008 MASE	Year 2008 SMAPE
STCD	1,430	0.640 ± 0.05	1.584 ± 0.11	0.199 ± 0.02	0.694 ± 0.04	1.353 ± 0.09	0.195 ± 0.03	0.618 ± 0.06	1.678 ± 0.13	0.294 ± 0.03
TCDF	1,331	0.780 ± 0.06	1.817 ± 0.14	0.216 ± 0.03	0.758 ± 0.05	1.463 ± 0.14	0.210 ± 0.03	0.694 ± 0.05	1.973 ± 0.14	0.314 ± 0.02
Selego	1,508	0.821 ± 0.10	1.892 ± 0.18	0.352 ± 0.03	1.830 ± 0.11	1.725 ± 0.13	0.310 ± 0.02	1.864 ± 0.06	2.237 ± 0.18	0.327 ± 0.02
All Features	3,129	0.875 ± 0.09	1.937 ± 0.14	0.375 ± 0.02	0.894 ± 0.11	1.818 ± 0.12	0.343 ± 0.02	0.842 ± 0.09	2.874 ± 0.21	0.374 ± 0.02
Random Features	1,518	1.370 ± 0.07	2.746 ± 0.14	0.226 ± 0.01	1.328 ± 0.10	1.712 ± 0.14	0.318 ± 0.02	1.267 ± 0.07	3.569 ± 0.19	0.379 ± 0.03
STREAMS	1,518	**0.519 ± 0.03**	**1.328 ± 0.08**	**0.121 ± 0.01**	**0.465 ± 0.03**	**1.012 ± 0.04**	**0.136 ± 0.01**	**0.427 ± 0.03**	**0.967 ± 0.06**	**0.149 ± 0.01**

Table 3. Evaluating the generalization performance over different years across three different metrics. The timespan of 1916-1926 was used to infer the causal graph for each model and the prediction is done for the years 2000, 2004, and 2008. For each metric, a lower value is preferred. The results in bold represent the best results (details of the experimental setup and the discussions are available at [165])

analysis. STCD addresses this gap by incorporating spatial information into the causal discovery process, enhancing accuracy in identifying crucial causal connections, as demonstrated in hydrological studies within the Texas river basin.

The need for more sophisticated spatio-temporal attention mechanisms is acknowledged to further the capabilities of spatio-temporal causal discovery algorithms.

Table 2 presents the outcomes of experiments testing the spatio-temporal Causal Discovery (STCD) method's ability to discern causal relationships vital for understanding hydrological processes, particularly in identifying key spatial locations for water flow prediction. Lacking a true causal graph for such physical processes, we base our evaluation on the premise that predictions using causal features should outperform those made with features from temporal causal discovery methods like TCDF [119], or when using all available features. This is grounded in the expectation that causal features lead to better prediction generalization [77]. The results confirm that STCD-selected spatio-causal features consistently enhance performance compared to non-spatio-causal features.

We note, however, that more research is needed in this space. For example, while STCD enforces spatial constraints in causal discovery through penalizing attention scores based on spatial distances, in a follow up work, we have shown that STREAMS [165], which leveraged a spatio-temporal autoencoder with reinforcement learning (RL) to learn the causal relationships (where the spatial component aids in following the spatial constraint and the RL component aids by optimizing the search space to learn the causal graph) can provide significant gains over the baseline STCD. Table 3 provides sample results. This illustrates the fact that there is significant room for research in effective spatio-causal discovery and its application to spatial inference and forecasting tasks.

5.3 Data Structures, Models, and Algorithms for spatio-temporal Causal Learning

Understanding the influence of causal ancestors on their descendants is vital for designing targeted interventions. For example, the impact of vaccination programs on disease transmission involves direct and indirect effects, influenced by various spatial variables such as population age, health, vaccine acceptance, and socio-economic factors.

Addressing spatio-temporal heterogeneity in policy interventions, [199] introduces a spatially interrupted time-series design to analyze mobility control policies during the COVID-19 pandemic. This approach elucidates the changing effects of policies over time and space, highlighting significant heterogeneities in causal impacts. Similarly, [32] proposes a causal framework tailored for spatio-temporal data, incorporating models for estimating causal effects and a hypothesis test for detecting overall causal relationships. This framework, leveraging an inverse probability weighting technique and a nonparametric approach for error estimation, facilitates causal analysis without distributional assumptions and accommodates potential confounders.

The authors offer a class of spatio-temporal stochastic process causal models that allow them to explicitly characterize and quantify the causative influence of a vector of covariates X on a real-valued response Y. They characterize the causal effect using a counterfactual paradigm, which includes comparing the result under two distinct scenarios: one with X set to a specified value and another with X set to some other value.

The proposed technique makes no distributional assumptions about the data-generation process and allows for the effect of an infinite number of latent confounders assuming these confounders do not alter over time. Further research is encouraged to enhance these methodologies, especially in integrating specific spatio-temporal assumptions and addressing confounders that vary over space and time.

As evidenced by the above discussion, studying spatio-temporal causality requires large amounts of spatio-temporal data (such as large transportation or river networks) and learning of spatio-causal models from observational data is expensive as identifying and assessing causality across space and time adds significant complexity to the causal learning process. Considering that, in the most general case, each distinct point in space can be considered as a separate causal variable, which may impact its neighbors and may, in turn, be impacted by them, the number of variables of interest in spatio-causal discovery can grow significantly, quickly outstriping the scalabilities of the causal learning algorithms. This necessitates research into effectively exploiting spatial data structures, network models, and multi-resolution algorithms within a spatio-causal learning framework to tackle the underlying efficiency and scalability challenges by effectively pruning and/or indexing the observations (and their inter-relationships) in time and space. This currently is a largely unexplored area of research, though results from early work on leveraging available spatial and/or causal metadata in improving efficiency, scalability, and accuracy challenges are promising [23, 99, 109, 148, 158, 182].

Due to the additional difficulty of determining causality in both dimensions, studying spatio-temporal causality requires large amounts of spatio-temporal data, such as from transportation or river networks. As a result, learning spatio-causal models from observational data is expensive.

As every point in space has the potential to be a distinct causal variable that influences and is influenced by its neighbors, the number of variables in spatio-causal discovery might grow exponentially, which puts the scalability of causal learning algorithms to the test. This scenario emphasizes the need for more study into how to manage the observations and their temporal and spatial linkages in a spatio-causal framework to better achieve efficiency and scalability through the use of multi-resolution algorithms, network models, and spatial data structures. Initial attempts to use spatial and causal metadata appear promising in improving accuracy, scalability, and efficiency, despite the fact that this is a new field of study [23, 99, 109, 148, 158, 182].

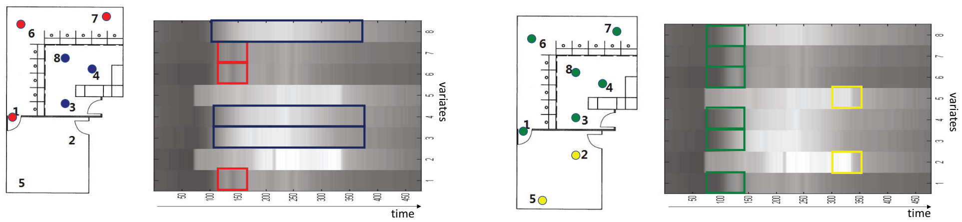


Fig. 9. Sample events identified on a spatio-temporal data set; here, each row corresponds to a different sensor – detected spatio-temporal events (with different temporal lengths and involving different groups of sensors) are highlighted with colored boxes on the time series; the sensors involved in these events are also recorded with colored dots on the spatial map. Note that the sensor readings are causally related to each other: similar changes in the same zone are likely caused by the same external cause; moreover, the spatial connectivity between the zones may imply that events can be spatially diffused and certain events (such as the ones marked with green color in the figure) can cover multiple spatial zones due to spatially diffused causal effects [102]

Validating spatio-causal models is further complicated by the need to assess the quality of acquired causal structures, especially in situations when a ground truth causal network is incomplete. In order to overcome this, the STCD approach uses generalization errors over time as a measure of accuracy to evaluate inferred causal linkages through time-series forecasting on the target variable. While evaluating spatio-temporal causal discovery algorithms is fundamentally hampered by the general lack of ground-truth data, this method evaluates the validity of found causal linkages in the absence of a ground-truth causal network.

6 SPATIO-TEMPORAL CAUSAL KNOWLEDGE FOR SPATIAL DATA STRUCTURES, MODELS, AND ALGORITHMS

Understanding spatio-causality is vital for managing and analyzing spatio-temporal data across various applications, such as point-of-interest detection, spatio-temporal event detection, and trajectory management. Recognizing the potential for advancement in this area, we identify key research opportunities including:

- Developing causally-aware multi-scale event extraction methods to enhance spatio-temporal search and analysis,
- Innovating causally-prioritized spatio-temporal data analysis techniques that reduce redundancy and improve resilience to noise and data sparsity, and
- Creating new causally informed indexing and search mechanisms.

This section will explore these challenges and the potential avenues for research.

6.1 Detecting and Classifying Events with Causal Awareness in Space and Time

Decisive choices often rely on the skill to identify and categorize important occurrences in complex, time and space-related information, with the goal of revealing hidden patterns that are not obvious at first glance. These questions are crucial:

- Identifying significant events in spatio-temporal datasets is a complex task that requires careful analysis and consideration.
- Can these events be categorized using current labels or by differentiating between typical and unusual activities?
- Is it possible to analyze spatio-temporal events to grasp their similarities and differences?
- Is there a way to analyze event patterns to reveal their connections?

Model	RMSE	MAE	MAPE
Bi-LSTM with no attention(74 epochs)	9.00	6.38	4.67
LSTM with no attention (74 epochs)	9.06	6.38	4.67
LSTM with single layer attention to input (20 epochs)	9.00	6.52	4.79
LSTM with single layer attention to output (20 epochs)	9.02	6.52	4.79
M2NN (15 epochs)	**8.63**	**6.24**	**4.63**

Table 4. Comparison of different architectures for a traffic prediction task; here M2NN is a multi-variate time series predictor which leverages RMT based spatio-temporal attention to reduce prediction errors (details of the experimental setup and the discussions are available at [148])

When dealing with spatio-temporal data, such as in water management and pandemic tracking (discussed in Sections 2.1 and 2.2), it is important to consider the specific spatio-causal features that are essential for analytical purposes (illustrated in Figure 9). Using metadata-aware techniques Strong and resilient Utilizing multi-variate Temporal (RMT) feature extraction methods, as discussed in [102, 189], can improve feature extraction accuracy and robustness by incorporating variate correlations. In particular, RMT features use metadata graphs to pinpoint dependency neighborhoods, analyzing various relationships to detect important local changes over time and space.

Recent advancements, such as the ones in [148, 182], have built upon RMT characteristics to predict intricate events in multi-dimensional time series, demonstrating their effectiveness in various spatio-temporal scenarios, such as COVID-19 forecasting and travel analysis (Table 4).

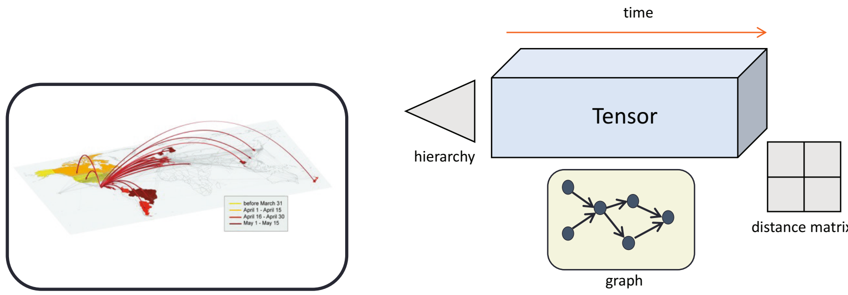
Various techniques for analyzing spatio-temporal events are investigated in [3, 42, 91, 101, 188], mainly using GNNs or spatio-temporal CNNs. These methods have a common approach to identifying events in spatio-temporal data through:

- Exploring connections between data variables to establish neighborhoods,
- Identifying unique patterns, recurring patterns, or specific target configurations.

In general, though, these tasks are complicated by the facts that (a) spatio-temporal data are often of mixed quality (and sometimes missing) due to the sensor variety and location, and measurement errors due to calibration difficulties. Moreover, as described above, (b) confounders and colliders can introduce spurious (non-causal) correlations in the data, which can result in noisy features/events that are not robust [29]. We therefore argue that spatio-causal knowledge can greatly improve spatio-temporal event detection and classification performance. In particular, spatio-causal knowledge can help improve both the effectiveness of the inter-relationships identified and leveraged across data and can help prioritize the patterns especially when given the sought after events are constrained towards a spatio-temporal target.

Careful thought is required when integrating causal information into spatio-temporal data analysis. As an example, in the RMT algorithm, neighborhoods are typically defined as follows: h-hop [33, 124]; reachability [18, 191, 196, 201]; cluster/partition [48, 83, 121]; and hitting distance neighborhoods [28, 111]. These definitions are based on proximity or strong coupling between graph nodes. To meet particular analysis demands, the idea of distance among nodes can be contextually adjusted [26, 78, 80, 87, 159, 184].

It is important to distinguish between direct causal impacts (such as the spread of a disease from one region to another) and indirect linkages mediated by confounders or intermediaries when focusing on causally-informed traits or events. In example, confounders can mask or dilute the true relationships in the data, which can cause errors in the extraction of features and events and reduce the usefulness of the features that are found for tasks like classification and similarity search.



(a) Sample epidemic scenario (b) spatio-temporal tensor

Fig. 10. (a) A sample epidemic scenario, with temporally evolving transmission paths and (b) a spatio-temporal tensor, where different modes represent data of temporal and/or spatial nature; in this paper, we further argue that there can additionally be causal connections between elements of the various modes, which may need to be taken into account for effective data analysis

This emphasizes the need for more study on spatio-causal neighborhood definitions and the advancement of spatio-causal convolution and smoothing methods. With the goal of providing a more precise and insightful analysis of spatio-temporal events, such techniques ought to strive to capture the complex spatio-topological and causal dynamics found in the data.

6.2 Causally-Prioritized spatio-temporal Analysis to Eliminate Redundant Work and Providing Robustness Against Noise and Sparsity

When applied to sparse, noisy, and inconsistent datasets, causal knowledge dramatically improves spatio-temporal analysis through increased resilience, decreased computations, and increased accuracy [163]. It applies causally-informed attention mechanisms, creates low-dimensional embeddings, and assists in feature selection. These contributions are essential for managing high-dimensional, complicated data, enabling dimensionality reduction, and enhancing the results of analyses in semi-supervised, supervised, and unsupervised contexts.

One excellent example of using causal knowledge is tensor analysis. It maps attributes to a multi-modal array enriched with network and spatial information to facilitate sophisticated spatio-temporal analysis and embeddings. It encodes multi-variate data over time as tensor streams (Figure 10). Using decomposition methods such as eigen-decomposition and Tucker decomposition, spectral features can be shown by breaking down tensors into factor matrices and a core matrix.

However, difficulties still exist and they are as follows:

- High computational costs and the intermediary data blow-up problem complicate the analysis of high-modal datasets, especially when data are sparse.
- Sensitivity to noisy data can lead to overfitting and erroneous conclusions, a significant concern in sparse web and social datasets.
- Recommendation tasks face inherent biases, such as popularity bias, affecting the fairness and relevance of outcomes.

To overcome these problems, we suggest adding spatio-causal knowledge, which provides a more sophisticated and useful method for spatio-temporal data analysis.

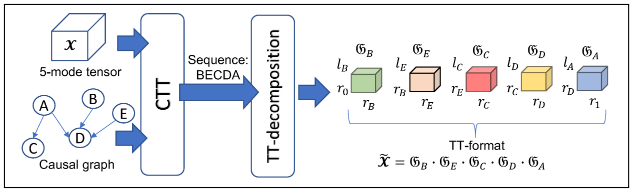
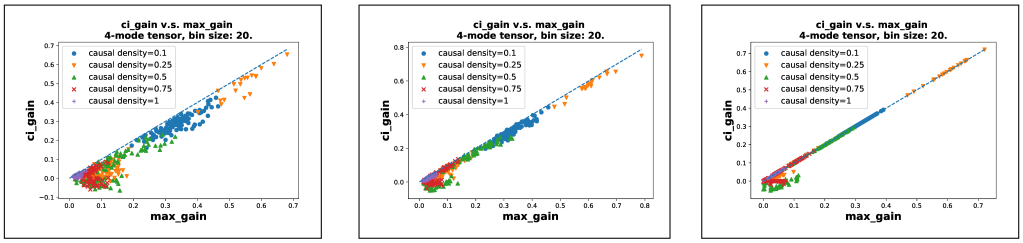


Fig. 11. Overview of the causally informed tensor train decomposition process



(a) TT-rank percentage = 50%

(b) TT-rank percentage = 75%

(c) TT-rank percentage= 100%

Fig. 12. ci_gain vs. max_gain for 100 sample graphs per configuration (4 mode tensors, each mode with 20 distinct values). One thing that emerges from these charts is that, especially for scenarios where the maximum possible gain is large, causally-informed tensor train decomposition (CTT) provides close to maximum gain and the performance is approaching to that of the optimal sequence when the TT-rank is sufficiently large (details of the experimental setup and the discussions are available at [109])

6.2.1 Managing Blowups of Intermediate Data. The curse of dimensionality impedes high-dimensional data analysis, particularly in multi-modal datasets represented as tensors. Tensor decomposition methods such as CP and Tucker attempt to embed data into lower-dimensional latent spaces in order to minimize dimensionality. Despite having a dense core that makes it information-rich, Tucker decomposition has exponential memory expansion as the number of tensor modes rises. In order to reduce both space and computing time, the Tensor Train (TT) decomposition divides the tensor into a number of smaller 3-modal cores. Optimizing the decomposition sequence is still a difficult task, though.

TT decomposition sequence is guided by data features such as entropy, as demonstrated in recent work [96]. However, as Pearl and colleagues [133] have shown, these approaches have limitations, and it is crucial to comprehend the underlying causal structures. Data analysis results might be unexpectedly impacted by the introduction or elimination of statistical dependencies during the conditioning process, which involves segmenting variable domains.

We suggest that the selection of decomposition sequences for TT can be improved by utilizing knowledge of the causal linkages among data modes, providing a more accurate and effective representation. As shown in (Figure 12), we may obtain optimal decomposition sequences by using algorithms that are guided by knowledge of the spatio-causal structure, which greatly enhances the analysis of high-dimensional data.

6.2.2 Improving Robustness. Our earlier work showed that by adapting to the noise profiles inside data [98], multi-resolution approaches can minimize time and memory utilization without affecting decomposition quality. In order to improve the performance of incremental updates for growing tensor streams and help with noise management, we proposed sub-tensor impact graphs (SIGs) to encode structural information of tensor sub-partitions [23]. SIGs show the relationship between decompositions and the spread of errors by directing the division of tensors through independent block decomposition and iterative integration of these decompositions. This method makes it easier to allocate resources (such as update schedules or decomposition rankings) in a way that maximizes robustness against sparsity and noise while balancing both accuracy and efficiency.

Dataset	Model	Popularity De-Biased Test Sets							
		Popularity = 2		Popularity = 3		Popularity = 5		Popularity = 10	
		NDCG@10	HR@10	NDCG@10	HR@10	NDCG@10	HR@10	NDCG@10	HR@10
Epinions	NeuMF [74]	0.30 ± 0.02	0.65 ± 0.02	0.30 ± 0.02	0.65 ± 0.02	0.31 ± 0.02	0.67 ± 0.01	0.33 ± 0.01	0.70 ± 0.01
	PMF [113]	0.32 ± 0.02	0.68 ± 0.02	0.33 ± 0.02	0.70 ± 0.01	0.36 ± 0.01	0.71 ± 0.02	0.39 ± 0.01	0.73 ± 0.01
	SocialMF [79]	0.31 ± 0.02	0.63 ± 0.02	0.33 ± 0.01	0.65 ± 0.02	0.34 ± 0.01	0.70 ± 0.01	0.36 ± 0.01	0.72 ± 0.01
	GraphRec [46]	0.23 ± 0.02	0.53 ± 0.02	0.23 ± 0.01	0.54 ± 0.02	0.20 ± 0.01	0.51 ± 0.01	0.22 ± 0.01	0.51 ± 0.02
	ConsisRec [197]	0.35 ± 0.02	0.70 ± 0.02	0.37 ± 0.01	0.74 ± 0.01	0.40 ± 0.02	0.76 ± 0.01	0.42 ± 0.01	0.78 ± 0.01
	IPS-MF [100]	0.30 ± 0.02	0.64 ± 0.03	0.31 ± 0.02	0.68 ± 0.01	0.36 ± 0.02	0.73 ± 0.01	0.40 ± 0.01	0.78 ± 0.01
	CIRS [190]	0.36 ± 0.03	0.69 ± 0.02	**0.39 ± 0.02**	0.73 ± 0.01	0.40 ± 0.01	0.73 ± 0.02	0.41 ± 0.01	0.80 ± 0.01
	DICE [200]	0.32 ± 0.02	0.67 ± 0.02	0.35 ± 0.01	0.70 ± 0.02	0.37 ± 0.01	0.73 ± 0.01	0.42 ± 0.01	0.82 ± 0.01
	D2Rec (ours)	**0.37 ± 0.02**	**0.73 ± 0.01**	0.38 ± 0.02	**0.74 ± 0.02**	**0.40 ± 0.01**	**0.81 ± 0.01**	**0.43 ± 0.02**	**0.83 ± 0.01**
Ciao	NeuMF [74]	0.24 ± 0.03	0.53 ± 0.02	0.26 ± 0.02	0.58 ± 0.02	0.30 ± 0.01	0.61 ± 0.01	0.32 ± 0.01	0.68 ± 0.02
	PMF [113]	0.29 ± 0.02	0.60 ± 0.01	0.32 ± 0.01	0.66 ± 0.02	0.36 ± 0.02	0.72 ± 0.03	0.41 ± 0.02	0.81 ± 0.01
	SocialMF [79]	0.32 ± 0.02	0.59 ± 0.02	0.34 ± 0.03	0.68 ± 0.02	0.37 ± 0.02	0.75 ± 0.03	0.42 ± 0.01	0.81 ± 0.02
	GraphRec [46]	0.17 ± 0.02	0.35 ± 0.02	0.13 ± 0.02	0.33 ± 0.03	0.14 ± 0.02	0.32 ± 0.03	0.15 ± 0.02	0.43 ± 0.01
	ConsisRec [197]	0.27 ± 0.03	0.56 ± 0.02	0.33 ± 0.02	0.61 ± 0.01	0.32 ± 0.02	0.66 ± 0.02	0.36 ± 0.01	0.71 ± 0.01
	IPS-MF [100]	0.24 ± 0.03	0.52 ± 0.03	0.31 ± 0.02	0.61 ± 0.02	0.39 ± 0.02	0.72 ± 0.01	0.42 ± 0.02	0.81 ± 0.01
	CIRS [190]	**0.34 ± 0.03**	**0.64 ± 0.03**	**0.35 ± 0.02**	0.66 ± 0.02	0.37 ± 0.01	0.71 ± 0.02	0.42 ± 0.01	0.80 ± 0.01
	DICE [200]	0.26 ± 0.03	0.55 ± 0.03	0.33 ± 0.02	0.64 ± 0.02	0.35 ± 0.01	0.70 ± 0.02	0.45 ± 0.01	0.81 ± 0.01
	D2Rec (ours)	0.33 ± 0.01	0.61 ± 0.02	0.34 ± 0.02	**0.68 ± 0.02**	**0.38 ± 0.03**	**0.78 ± 0.02**	**0.48 ± 0.02**	**0.88 ± 0.01**

Table 5. Comparing the ranking performance of different recommender system models with D2Rec for *Epinions* and *Ciao* across 10 runs over popularity de-biased data sets (since item popularity is an aspect of confounding, by generating popularity de-biased test sets and measuring D2Rec's performance over them, we can verify whether D2Rec effectively adjusts for the confounding bias). Results in bold represent the best results and the ones underlined represent the second best; note that causally informed D2Rec consistently outperforms the competitors (details of the experimental setup and the discussions are available at [162]) – the challenge, of course, is to build on these results in a context where causal networks are superimposed on spatial networks

Moreover, variate interdependencies provide valuable insights for variate-selection and attention tactics that improve the resilience of forecasting algorithms. For example, because of their higher noise resilience, simpler, metadata-supported models like CNNs can forecast more accurately than more sophisticated models like LSTMs [148, 182].

Accurately recognizing and measuring these interdependencies is the difficult part. We suggest that more accurate and reliable results will result from decomposition strategies that incorporate spatial causality knowledge. Tensor "slices" are often analyzed via decomposition algorithms in order to gradually improve the overall decomposition. However, the underlying causal structure affects how well these slice-based conditioning processes work and might induce errors in analysis or hide important information. In order to ensure that decompositions are extremely accurate and robust to sparsity and noise, we thereby support causally informed slicing and partitioning of spatio-temporal data, highlighting the need for additional research into causally informed tensor decomposition algorithms.

6.2.3 Improving Bias Resilience . Recommender systems, trained on user-item interactions, often grapple with biases like selection and popularity bias due to the missing-not-at-random nature of their observational training data. The non-random exposure mechanism, which is determined by the spatio-temporal context of the item and the interests of the user, intensifies this bias. As a result, these algorithms may overemphasize well-liked products while ignoring user-specific interests and creating feedback loops that reinforce prejudices.

Re-weighting techniques have been the mainstay of previous attempts to reduce these biases [100]. A more sophisticated strategy, on the other hand, proposes modeling the item exposure mechanism to users and moving toward a causal framework for suggestions. This strategy is still untested despite its potential.

Our previous work filled this gap by introducing techniques such as the Social Implicit-Disentangled Recommender System (SIDR [161]) and the Disentangled and De-Biased Recommender (D2Rec [162]), which use auxiliary network information to improve de-biasing accuracy by applying causal inference techniques to disentangle factors influencing item exposure, rating predictions, and confounders.

To produce fully de-biased suggestions, however, the integration of spatio-temporal contexts and spatio-causal structures into recommender systems remains unexplored. This points to a major area of untapped potential for future research: recommender systems capable of navigating the intricacies of spatio-temporal data and providing objective, user-specific item recommendations.

6.3 Causally-Informed Indexing and Search

Spatial-causal information can be used to improve the recognition and ranking of spatio-temporal features and events. It can also be used to create robust, efficient low-dimensional embeddings for spatio-temporal data. Using this knowledge to improve indexing and searching operations is one immediate application.

A vast amount of research has been done on exact and approximate spatial (2D, 3D) and multi-dimensional data structures [155, 156]. These include tree-based data structures [12, 15, 53, 70, 160], and spatial hashing algorithms [94], as well as grid-based and Voronoi-diagram based methods [154]. While these structures are capable of handling a variety of search operations, they frequently face difficulties with high query latency and the real-time upkeep of indices for moving objects.

According to this paper, spatio-causal information can address these issues in two important ways:

Causally-informed characterization of query workload: Enhancing throughput is recognized to be possible through index structure and search process optimization to handle query workloads [57, 195]. With its sophisticated comprehension of data dynamics, spatio-causal information provides a novel way to monitor and adjust to changes in query workloads more successfully. In addition to workload adaptation, index structures could be created or adjusted to match underlying causal structures, possibly by means of data/space partitions or embeddings that are causally informed, as covered in Section 6.2. This method may also help search methods by directing the order in which partitions are investigated according to spatio-causal correlations.

Characterizing updates with causal information: Updating index structures in real-time can be a difficult task, especially when dealing with concurrent queries. An understanding of the causality of spatial processes such as movement patterns may improve update scheduling and update policy formulation. The responsiveness and efficiency of the data structures can be enhanced by implementing techniques for both eager and lazy updates more skillfully by knowing the underlying causes of data changes.

7 CONCLUSIONS

In this paper, we have made a case for urgent effort into "spatio-causal" research, including research into (a) spatio-causal discovery and inference, (b) causally-informed spatial data structures, models, and algorithms, and (c) new spatial data structures, models, and algorithms to support efficient spatio-temporal causal learning. We have motivated our vision with two urgent applications in sustainability and public health and outlined how these two applications, along with many other socio-economically critical, human-centered application domains that share common spatio-temporal challenges with these two, can benefit from the proposed paradigm of spatio-causal research.

REFERENCES

[1] HAWQS 2.0. 2023. HAWQS System 2.0 and Data to model the lower 48 conterminous U.S using the SWAT model. https://doi.org/10.18738/T8/GDOPBA, Texas Data Repository, V2.

[2] Ibrahim Abubakar, Philippe Gautret, Gary W Brunette, Lucille Blumberg, David Johnson, Gilles Poumerol, Ziad A Memish, Maurizio Barbeschi, and Ali S Khan. 2012. Global perspectives for prevention of infectious diseases associated with mass gatherings. *The Lancet infectious diseases* 12, 1 (2012), 66–74.

[3] IY Agarwal, DP Rana, M Shaikh, and S Poudel. 2022. Spatio-temporal approach for classification of COVID-19 pandemic fake news. *Social Network Analysis and Mining* 12, 1 (2022), 68.

[4] Angelos Alamanos, Alec Rolston, and George Papaioannou. 2021. Development of a decision support system for sustainable environmental management and stakeholder engagement. *Hydrology* 8, 1 (2021), 40.

[5] Michael Anderson and Jeremy Magruder. 2012. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal* 122, 563 (2012), 957–989.

[6] Roy M Anderson and Robert M May. 1991. *Infectious diseases of humans: dynamics and control*. Oxford university press.

[7] Sinan Aral and Christos Nicolaides. 2017. Exercise contagion in a global social network. *Nature communications* 8, 1 (2017), 14753.

[8] Andrew Arnold, Yan Liu, and Naoki Abe. 2007. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 66–75.

[9] Fahim Tasneema Azad, Robert W Dodge, Allen M Varghese, Jaejin Lee, Giulia Pedrielli, K Selçuk Candan, and Gerardo Chowell-Puente. 2022. Sirtem: Spatially informed rapid testing for epidemic modeling and response to covid-19. *ACM Transactions on Spatial Algorithms and Systems* 8, 4 (2022), 1–43.

[10] Fahim Tasneema Azad, Javier Redondo Antón, Bilgehan Arslan, K. Selçuk Candan, Maria Luisa Sapino, Gerardo Chowell-Puente, Giulia Pedrielli, Shubhodeep Mitra, Fateh Singh, Hans Behrens, and Mao-Lin Li. 2023. Pancommunity: Non-Monolithic Complex Epidemic and Pandemic Modeling. In *9th International Conference on Infectious Disease Dynamics (P3.172)*. Available at SSRN: https://ssrn.com/abstract=4655113 or http://dx.doi.org/10.2139/ssrn.4655113.

[11] Elias Bareinboim, Jin Tian, and Judea Pearl. 2022. Recovering from selection bias in causal and statistical inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. 433–450.

[12] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. 1990. The R*-tree: An efficient and robust access method for points and rectangles. In *Proceedings of the 1990 ACM SIGMOD international conference on Management of data*. 322–331.

[13] Hans Walter Behrens, K Selçuk Candan, Xilun Chen, Ashish Gadkari, Yash Garg, Mao-Lin Li, Xinsheng Li, Sicong Liu, Nicholas Martinez, Jiayong Mo, Elliot Nester, Silvestro Poccia, Manjusha Ravindranath, and Maria Luisa Sapino. 2018. Datastorm-FE: A data-and decision-flow and coordination engine for coupled simulation ensembles. *Proceedings of the VLDB Endowment* 11, 12 (2018), 1906–1909.

[14] Hans Walter Behrens, K Selçuk Candan, Xilun Chen, Yash Garg, Mao-Lin Li, Xinsheng Li, Sicong Liu, Maria Luisa Sapino, Md Shadab, Dalton Turner, and Magesh Vijayakumaren. 2021. DataStorm: Coupled, Continuous Simulations for Complex Urban Environments. *ACM/IMS Transactions on Data Science* 2, 3 (2021), 1–37.

[15] Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* 18, 9 (1975), 509–517.

[16] Joseph Berkson. 1946. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* 2, 3 (1946), 47–53.

[17] Alexander Bochman. 2018. Actual causality in a logical setting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 1730–1736.

[18] Paolo Boldi, Marco Rosa, and Sebastiano Vigna. 2011. HyperANF: Approximating the Neighbourhood Function of Very Large Graphs on a Budget. In *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad, India) *(WWW '11)*. Association for Computing Machinery, New York, NY, USA, 625–634. https://doi.org/10.1145/1963405.1963493

[19] Kenneth Bollen and Judea Pearl. 2013. *Eight Myths About Causality and Structural Equation Models*. 301–328. https://doi.org/10.1007/978-94-007-6094-3_15

[20] Kenneth A Bollen and Judea Pearl. 2013. Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*. Springer, 301–328.

[21] Greg Browder, Suzanne Ozment, Irene Rehberger Bescos, Todd Gartner, and Glenn-Marie Lange. 2019. *Integrating green and gray*. Washington, DC: World Bank and World Resources Institute.

[22] Donald S Burke, Joshua M Epstein, Derek AT Cummings, Jon I Parker, Kenneth C Cline, Ramesh M Singa, and Shubha Chakravarty. 2006. Individual-based computational modeling of smallpox epidemic control strategies. *Academic Emergency Medicine* 13, 11 (2006), 1142–1149.

[23] K Selçuk Candan, Shengyu Huang, Xinsheng Li, and Maria Luisa Sapino. 2019. *Clustering Methods for Big Data Analytics: Techniques, Toolboxes and Applications*. Springer, Chapter Effective Tensor-Based Data Clustering Through Sub-Tensor Impact Graphs, 145–179.

[24] Christopher Carpenter and Carlos Dobkin. 2009. The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics* 1, 1 (2009), 164–182.

[25] Simon Cauchemez, Neil M Ferguson, Claude Wachtel, Anders Tegnell, Guillaume Saour, Ben Duncan, and Angus Nicoll. 2009. Closure of schools during an influenza pandemic. *The Lancet infectious diseases* 9, 8 (2009), 473–481.

[26] Soumen Chakrabarti. 2007. Dynamic personalized pagerank in entity-relation graphs. In *Proceedings of the 16th international conference on World Wide Web*. 571–580.

[27] Dennis L Chao, M Elizabeth Halloran, Valerie J Obenchain, and Ira M Longini Jr. 2010. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS computational biology* 6, 1 (2010), e1000656.

[28] Mo Chen, Jianzhuang Liu, and Xiaoou Tang. 2008. Clustering via Random Walk Hitting Time on Directed Graphs. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2* (Chicago, Illinois) *(AAAI'08)*. AAAI Press, 616–621.

[29] Lu Cheng, Ruocheng Guo, Kasim Selçuk Candan, and Huan Liu. 2022. Effects of Multi-Aspect Online Reviews with Unobserved Confounders: Estimation and Implication. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, Ceren Budak, Meeyoung Cha, and Daniele Quercia (Eds.). AAAI Press, 67–78. https://ojs.aaai.org/index.php/ICWSM/article/view/19273

[30] Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, K. Selçuk Candan, and Huan Liu. 2022. Evaluation Methods and Measures for Causal Learning Algorithms. *IEEE Trans. Artif. Intell.* 3, 6 (2022), 924–943. https://doi.org/10.1109/TAI.2022.3150264

[31] Gerardo Chowell and Cécile Viboud. 2016. Is it growing exponentially fast?–impact of assuming exponential growth for characterizing and forecasting epidemics with initial near-exponential growth dynamics. *Infectious disease modelling* 1, 1 (2016), 71–78.

[32] Rune Christiansen, Matthias Baumann, Tobias Kuemmerle, Miguel D Mahecha, and Jonas Peters. 2022. Toward causal inference for spatio-temporal data: conflict and forest loss in colombia. *J. Amer. Statist. Assoc.* 117, 538 (2022), 591–601.

[33] Edith Cohen, Eran Halperin, Haim Kaplan, and Uri Zwick. 2002. Reachability and Distance Queries via 2-Hop Labels. In *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (San Francisco, California) *(SODA '02)*. Society for Industrial and Applied Mathematics, USA, 937–946.

[34] Thomas D Cook, Donald Thomas Campbell, and William Shadish. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Vol. 1195. Houghton Mifflin Boston, MA.

[35] Kyran Cupido, Petar Jevtić, and Antonio Paez. 2020. Spatial patterns of mortality in the United States: A spatial filtering approach. *Insurance: Mathematics and Economics* 95 (2020), 28–38.

[36] Thomas E Dahl. 2011. *Status and trends of wetlands in the conterminous United States 2004 to 2009*. US Department of the Interior, US Fish and Wildlife Service, Fisheries and ....

[37] Hamidreza Ghasemi Damavandi, Reepal Shah, Dimitrios Stampoulis, Yuhang Wei, Dragan Boscovic, and John Sabo. 2019. Accurate prediction of streamflow using long short-term memory network: a case study in the Brazos River Basin in Texas. *International Journal of Environmental Science and Development* 10, 10 (2019), 294–300.

[38] Hamidreza Ghasemi Damavandi, Dimitrios Stampoulis, Reepal Shah, Yuhang Wei, Dragan Boscovic, and John Sabo. 2019. Machine learning: an efficient alternative to the variable infiltration capacity model for an accurate simulation of runoff rates. *International Journal of Environmental Science and Development* 10, 9 (2019), 288–293.

[39] Thomas K Dasaklis, Costas P Pappis, and Nikolaos P Rachaniotis. 2012. Epidemics control and logistics operations: A review. *International Journal of Production Economics* 139, 2 (2012), 393–410.

[40] Peter Daszak, Andrew A Cunningham, and Alex D Hyatt. 2001. Anthropogenic environmental change and the emergence of infectious diseases in wildlife. *Acta tropica* 78, 2 (2001), 103–116.

[41] Odo Diekmann and Johan Andre Peter Heesterbeek. 2000. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Vol. 5. John Wiley & Sons.

[42] Ziluo Ding, Rui Zhao, Jiyuan Zhang, Tianxiao Gao, Ruiqin Xiong, Zhaofei Yu, and Tiejun Huang. 2022. Spatio-temporal recurrent networks for event-based optical flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 525–533.

[43] Otis Dudley Duncan. 1966. Path analysis: Sociological examples. *American journal of Sociology* 72, 1 (1966), 1–16.

[44] Andrew C Eggers, Ronny Freier, Veronica Grembi, and Tommaso Nannicini. 2018. Regression discontinuity designs based on population thresholds: Pitfalls and solutions. *American Journal of Political Science* 62, 1 (2018), 210–229.

[45] Kevin L Erwin. 2009. Wetlands and global climate change: the role of wetland restoration in a changing world. *Wetlands Ecology and management* 17, 1 (2009), 71–84.

[46] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The World Wide Web Conference*. 417–426.

[47] Nina H. Fefferman, John S. McAlister, Belinda S. Akpa, Kelechi Akwataghibe, Fahim Tasneema Azad, Katherine Barkley, Amanda Bleichrodt, Michael J. Blum, L. Bourouiba, Yana Bromberg, K. Selçuk Candan, Gerardo Chowell, Erin Clancey, Fawn A. Cothran, Sharon N. DeWitte, Pilar Fernandez, David Finnoff, D. T. Flaherty, Nathaniel L. Gibson, Natalie Harris, Qiang He, Eric T. Lofgren, Debra L. Miller, James Moody, Kaitlin Muccio, Charles L. Nunn, Monica Papeș, Ioannis Ch. Paschalidis, Dana K. Pasquale, J. Michael Reed, Matthew B. Rogers, Courtney L. Schreiner, Elizabeth B. Strand, Clifford S. Swanson, Heather L. Szabo-Rogers, and Sadie J. Ryan. 2023. A New Paradigm for Pandemic Preparedness. *Curr Epidemiol Rep* 10 (2023), 240–251. https://doi.org/10.1007/s40471-023-00336-w

[48] Uriel Feige, MohammadTaghi Hajiaghayi, and James R. Lee. 2008. Improved Approximation Algorithms for Minimum Weight Vertex Separators. *SIAM J. Comput.* 38, 2 (2008), 629–657.

[49] Eli P Fenichel, Carlos Castillo-Chavez, M Graziano Ceddia, Gerardo Chowell, Paula A Gonzalez Parra, Graham J Hickling, Garth Holloway, Richard Horan, Benjamin Morin, Charles Perrings, Michael Springborn, Leticia Velazquez, and Cristina Villalobos. 2011. Adaptive human behavior in epidemiological models. *Proceedings of the National Academy of Sciences* 108, 15 (2011), 6306–6311.

[50] Eli P Fenichel, Nicolai V Kuminoff, and Gerardo Chowell. 2013. Skip the trip: Air Travelers' behavioral responses to pandemic influenza. *PloS one* 8, 3 (2013), e58249.

[51] Neil M Ferguson, Derek AT Cummings, Simon Cauchemez, Christophe Fraser, Steven Riley, Aronrag Meeyai, Sopon Iamsirithaworn, and Donald S Burke. 2005. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437, 7056 (2005), 209–214.

[52] Neil M Ferguson, Derek AT Cummings, Christophe Fraser, James C Cajka, Philip C Cooley, and Donald S Burke. 2006. Strategies for mitigating an influenza pandemic. *Nature* 442, 7101 (2006), 448–452.

[53] Raphael A Finkel and Jon Louis Bentley. 1974. Quad trees a data structure for retrieval on composite keys. *Acta informatica* 4 (1974), 1–9.

[54] Ronald A Fisher. 1922. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 222, 594-604 (1922), 309–368.

[55] Centers for Disease Control, Prevention, et al. 2007. Interim pre-pandemic planning guidance: community strategy for pandemic influenza mitigation in the United States-early, targeted, layered use of nonpharmaceutical interventions. *http://www. pandemicflu. gov/plan/community/community_mitigation. pdf* (2007).

[56] David A Freedman. 1987. As others see us: A case study in path analysis. *Journal of educational statistics* 12, 2 (1987), 101–128.

[57] Abdullah Gani, Aisha Siddiqa, Shahaboddin Shamshirband, and Fariza Hanum. 2016. A survey on indexing techniques for big data: taxonomy and performance evaluation. *Knowledge and information systems* 46 (2016), 241–284.

[58] Yash Garg and K Selçuk Candan. 2019. Racknet: Robust allocation of convolutional kernels in neural networks for image classification. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 315–323.

[59] Yash Garg and K Selçuk Candan. 2020. iSparse: Output informed sparsification of neural network. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 180–188.

[60] Yash Garg and K Selçuk Candan. 2021. Sdma: Saliency-driven mutual cross attention for multi-variate time series. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 7242–7249.

[61] Yash Garg, K Selçuk Candan, and Maria Luisa Sapino. 2020. San: Scale-space attention networks. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 853–864.

[62] Brian J Gerber. 2007. Disaster management in the United States: Examining key political and policy challenges. *Policy Studies Journal* 35, 2 (2007), 227–238.

[63] Hamidreza Ghasemi Damavandi, Dimitrios Stampoulis, John Sabo, Reepal Shah, Li Huang, Yuhang Wei, Yushiou Tsai, Jaishri Srinivasan, Tushar Sinha, Dragan Boscovic, and Glen Low. 2020. A Bayesian Neural Network for an Accurate Representation and Transformation of Runoff Dynamics: A Case Study of the Brazos River Basin in Texas. (02 2020). https://doi.org/10.12974/2311-8741.2020.08.5

[64] Christian Gische, Stephen G West, and Manuel C Voelkle. 2021. Forecasting causal effects of interventions versus predicting future outcomes. *Structural Equation Modeling: A Multidisciplinary Journal* 28, 3 (2021), 475–492.

[65] Robert J Glass, Laura M Glass, Walter E Beyeler, and H Jason Min. 2006. Targeted social distancing designs for pandemic influenza. *Emerging infectious diseases* 12, 11 (2006), 1671.

[66] M Maria Glymour and Sander Greenland. 2008. Causal diagrams. *Modern epidemiology* 3 (2008), 183–209.

[67] Adam N Glynn and Konstantin Kashin. 2018. Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1040–1049.

[68] Nicole L Gottdenker, Daniel G Streicker, Christina L Faust, and CR Carroll. 2014. Anthropogenic land use change and infectious diseases: a review of the evidence. *EcoHealth* 11 (2014), 619–632.

[69] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–37.

[70] Antonin Guttman. 1984. R-trees: A dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*. 47–57.

[71] Ned Hall. 2007. Structural equations and causation. *Philosophical Studies* 132 (2007), 109–136.

[72] Ned Hall, John Collins, and Laurie Paul. 2004. *Two concepts of causation.* MIT Press, Cambridge, MA.

[73] Leslie Hayduk, Greta Cummings, Rainer Stratkotter, Melanie Nimmo, Kostyantyn Grygoryev, Donna Dosman, Michael Gillespie, Hannah Pazderka-Robinson, and Kwame Boadu. 2003. Pearl's D-separation: One more step into causal thinking. *Structural Equation Modeling* 10, 2 (2003), 289–311.

[74] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[75] David Heckerman, Christopher Meek, and Gregory Cooper. 2006. A Bayesian approach to causal discovery. *Innovations in Machine Learning: Theory and Applications* (2006), 1–28.

[76] Rick H Hoyle. 2012. *Handbook of structural equation modeling.* Guilford press.

[77] Yeping Hu, Xiaogang Jia, Masayoshi Tomizuka, and Wei Zhan. 2022. Causal-based Time Series Domain Generalization for Vehicle Intention Prediction. In *ICRA*.

[78] Shengyu Huang, Xinsheng Li, K Selçuk Candan, and Maria Luisa Sapino. 2014. "Can you really trust that seed?": Reducing the impact of seed noise in personalized PageRank. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE, 216–223.

[79] Mohsen Jamali and Martin Ester. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*. 135–142.

[80] Glen Jeh and Jennifer Widom. 2003. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*. 271–279.

[81] Kurt Jensen, Lars Michael Kristensen, and Lisa Wells. 2007. Coloured Petri Nets and CPN Tools for modelling and validation of concurrent systems. *International Journal on Software Tools for Technology Transfer* 9 (2007), 213–254.

[82] Petar Jevtić and Luca Regis. 2019. A continuous-time stochastic model for the mortality surface of multiple populations. *Insurance: Mathematics and Economics* 88 (2019), 181–195.

[83] George Karypis and Vipin Kumar. 1998. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM Journal on Scientific Computing* 20, 1 (1998), 359–392.

[84] Kamran Khan, Julien Arino, Wei Hu, Paulo Raposo, Jennifer Sears, Felipe Calderon, Christine Heidebrecht, Michael Macdonald, Jessica Liauw, Angie Chan, et al. 2009. Spread of a novel influenza A (H1N1) virus via global airline transportation. *New England journal of medicine* 361, 2 (2009), 212–214.

[85] A Marm Kilpatrick and Sarah E Randolph. 2012. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. *The Lancet* 380, 9857 (2012), 1946–1955.

[86] JinHyung Kim and Judea Pearl. 1983. A computational model for causal and diagnostic reasoning in inference systems. In *International Joint Conference on Artificial Intelligence*. 0–0.

[87] Jung Hyun Kim, K Selçuk Candan, and Maria Luisa Sapino. 2013. Lr-ppr: Locality-sensitive, re-use promoting, approximate personalized pagerank computation. In *Proceedings of the 22nd ACM international conference on information & knowledge management*. 1801–1806.

[88] Keith W Kintigh, Katherine A Spielmann, Adam Brin, K Selçuk Candan, Tiffany C Clark, and Matthew Peeples. 2018. Data integration in the service of synthetic research. *Advances in Archaeological Practice* 6, 1 (2018), 30–41.

[89] Isidore K Kouadio, Syed Aljunid, Taro Kamigaki, Karen Hammad, and Hitoshi Oshitani. 2012. Infectious diseases following natural disasters: prevention and control measures. *Expert review of anti-infective therapy* 10, 1 (2012), 95–104.

[90] Frederik Kratzert, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger. 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences* 22, 11 (2018), 6005–6022.

[91] Parthasarathy Kulithalai Shiyam Sundar and Paresh Chandra Deka. 2022. Spatio-temporal classification and prediction of land use and land cover change for the Vembanad Lake system, Kerala: a machine learning approach. *Environmental Science and Pollution Research* 29, 57 (2022), 86220–86236.

[92] Kevin D Lafferty. 2009. The ecology of climate change and infectious diseases. *Ecology* 90, 4 (2009), 888–900.

[93] Eric F Lambin, Annelise Tran, Sophie O Vanwambeke, Catherine Linard, and Valérie Soti. 2010. Pathogenic landscapes: interactions between land, people, disease vectors, and their animal hosts. *International journal of health geographics* 9, 1 (2010), 1–13.

[94] Sylvain Lefebvre and Hugues Hoppe. 2006. Perfect spatial hashing. *ACM Transactions on Graphics (TOG)* 25, 3 (2006), 579–588.

[95] David Lewis. 1986. Postscripts to "Causation". In *Philosophical Papers Vol. Ii*, David Lewis (Ed.). Oxford University Press.

[96] Mao-Lin Li, K Selçuk Candan, and Maria Luisa Sapino. 2020. GTT: guiding the tensor train decomposition. In *Similarity Search and Applications: 13th International Conference, SISAP 2020, Copenhagen, Denmark, September 30–October 2, 2020, Proceedings*. Springer, 187–202.

[97] Mao-Lin Li, Francesco Di Mauro, K Selcuk Candan, and Maria Luisa Sapino. 2019. Matrix factorization with interval-valued data. *IEEE Transactions on Knowledge and Data Engineering* 33, 4 (2019), 1644–1658.

[98] Xinsheng Li, K Selçuk Candan, and Maria Luisa Sapino. 2017. nTD: noise-profile adaptive tensor decomposition. In *Proceedings of the 26th International Conference on World Wide Web*. 243–252.

[99] Xinsheng Li, Shengyu Huang, Kasim Selçuk Candan, and Maria Luisa Sapino. 2014. Focusing Decomposition Accuracy by Personalizing Tensor Decomposition (PTD). In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, Jianzhong Li, Xiaoyang Sean Wang, Minos N. Garofalakis, Ian Soboroff, Torsten Suel, and Min Wang (Eds.). ACM, 689–698. https://doi.org/10.1145/2661829.2662051

[100] Dawen Liang, Laurent Charlin, and David M Blei. 2016. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI. AUAI*.

[101] Yuan Liang, James Caverlee, and Cheng Cao. 2015. A noise-filtering approach for spatio-temporal event detection in social media. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings 37*. Springer International Publishing, 233–244.

[102] Sicong Liu, Silvestro Roberto Poccia, K Selcuk Candan, Maria Luisa Sapino, and Xiaolan Wang. 2018. Robust multi-variate temporal features of multi-variate time series. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 1 (2018), 1–24.

[103] Ting Liu, Qi Deng, Kaize Ding, John L. Sabo, Huan Liu, K. Selçuk Candan, and Rebecca Logsdon Muenich. 2023. Applying Graph Neural Networks to Improve the Data Resolution of Stream Water Quality Monitoring Networks. In *American Geophysical Union (AGU) 2023*.

[104] Aurélie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. 2009. Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics* 25, 12 (2009), i110–i118.

[105] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 23, 19 (2004), 2937–2960.

[106] Pratanu Mandal, Bilgehan Arslan, Paras Sheth, Rebecca Muenich, and K. Selcçuk Candan. 2023. Streamflow Prediction using SpatioTemporal Deep-Learning based Framework with Reinforcement Learning. In *HydroML'23 Symposium (poster)*.

[107] P. Mandal, Y. Choi, R. Shah, J. Sabo, H. Liu, and K. Selçuk Candan. 2024. Identifying Potential Wetlands via Causality-based Data Amputation and Knowledge Transfer. In *N-EWN Symposium*.

[108] Subramani Mani and Gregory F Cooper. 2000. Causal discovery from medical textual data.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 542.

[109] K. Selçuk Candan Mao-Lin Li and Maria Luisa Sapino. 2023. CTT: Causally Informed Tensor Train Decomposition. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 1180–1187.

[110] Rosemary A McFarlane, Adrian C Sleigh, and Anthony J McMichael. 2013. Land-use change and emerging infectious disease on an island continent. *International Journal of Environmental Research and Public Health* 10, 7 (2013), 2699–2719.

[111] Qiaozhu Mei, Denny Zhou, and Kenneth Church. 2008. Query Suggestion Using Hitting Time. In *CIKM '08 Proceedings of the 17th ACM conference on Information and knowledge management* (cikm '08 proceedings of the 17th acm conference on information and knowledge management ed.). ACM Press, 469–478. https://www.microsoft.com/en-us/research/publication/query-suggestion-using-hitting-time/

[112] George J Milne, Joel K Kelso, Heath A Kelly, Simon T Huband, and Jodie McVernon. 2008. A small community model for the transmission of infectious diseases: comparison of school closure as an intervention in individual-based models of an influenza pandemic. *PloS one* 3, 12 (2008), e4005.

[113] Andriy Mnih and Russ R Salakhutdinov. 2008. Probabilistic matrix factorization. In *Advances in neural information processing systems*. 1257–1264.

[114] Karthika Mohan, Judea Pearl, and Jin Tian. 2013. Graphical models for inference with missing data. *Advances in neural information processing systems* 26 (2013).

[115] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research* 17, 1 (2016), 1103–1204.

[116] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (2020), 18–33.

[117] Christo Morison, Małgorzata Fic, Thomas Marcou, Javad Mohamadichamgavi, Javier Redondo Antón, Golsa Sayyar, Alexander Stein, Frank Bastian, Hana Krakovska, Nandakishor Krishnan, Diogo Pires, Mohammad Reza Satouri, Frederik Jasper Thomsen, Kausutua Tjikundi, and Wajid Ali. 2024. Public Goods Games in Disease Evolution and Spread. https://doi.org/10.5281/ZENODO.10719143

[118] Mike Muller, Asit Biswas, Roberto Martin-Hurtado, and Cecilia Tortajada. 2015. Built infrastructure is essential. *Science* 349, 6248 (2015), 585–586.

[119] Meike Nauta, Doina Bucur, and Christin Seifert. 2019. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction* (2019).

[120] Grey S Nearing, Frederik Kratzert, Alden Keefe Sampson, Craig S Pelissier, Daniel Klotz, Jonathan M Frame, Cristina Prieto, and Hoshin V Gupta. 2021. What role does hydrological science play in the age of machine learning? *Water Resources Research* 57, 3 (2021), e2020WR028091.

[121] M. E. J. Newman. 2006. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74 (Sep 2006), 036104. Issue 3. https://doi.org/10.1103/PhysRevE.74.036104

[122] Hiroshi Nishiura, Carlos Castillo-Chavez, Muntaser Safan, and Gerardo Chowell. 2009. Transmission potential of the new influenza A (H1N1) virus and its age-specificity in Japan. *Eurosurveillance* 14, 22 (2009), 19227.

[123] World Health Organization et al. 2020. Coronavirus disease (COVID-19). (2020).

[124] Christopher R. Palmer, Phillip B. Gibbons, and Christos Faloutsos. 2002. ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Edmonton, Alberta, Canada) *(KDD '02)*. Association for Computing Machinery, New York, NY, USA, 81–90. https://doi.org/10.1145/775047.775059

[125] Margaret A Palmer, Junguo Liu, John H Matthews, Musonda Mumba, and Paolo D'Odorico. 2015. Manage water in a green way. *Science* 349, 6248 (2015), 584–585.

[126] Jonathan A Patz, Peter Daszak, Gary M Tabor, A Alonso Aguirre, Mary Pearl, Jon Epstein, Nathan D Wolfe, A Marm Kilpatrick, Johannes Foufopoulos, David Molyneux, and David Bradley. 2004. Unhealthy landscapes: policy recommendations on land use change and infectious disease emergence. *Environmental health perspectives* 112, 10 (2004), 1092–1098.

[127] Boris I Pavlin, Lisa M Schloegel, and Peter Daszak. 2009. Risk of importing zoonotic diseases through wildlife trade, United States. *Emerging infectious diseases* 15, 11 (2009), 1721.

[128] Judea Pearl. 1985. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*. 15–17.

[129] Judea Pearl. 1994. A probabilistic calculus of actions. In *Uncertainty Proceedings 1994*. Elsevier, 454–462.

[130] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

[131] Judea Pearl. 1998. Graphs, causality, and structural equation models. *Sociological Methods & Research* 27, 2 (1998), 226–284.

[132] Judea Pearl. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* 19, 2 (2000), 3.

[133] Judea Pearl. 2009. Causal inference in statistics: An overview. (2009).

[134] Judea Pearl. 2009. *Causality*. Cambridge university press.

[135] Judea Pearl. 2010. Causal inference. *Causality: objectives and assessment* (2010), 39–58.

[136] Judea Pearl. 2011. Bayesian networks. (2011).

[137] Judea Pearl. 2022. Fusion, propagation, and structuring in belief networks. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. 139–188.

[138] Judea Pearl and Elias Bareinboim. 2022. External validity: From do-calculus to transportability across populations. In *Probabilistic and causal inference: The works of Judea Pearl*. 451–482.

[139] Judea Pearl and Thomas S Verma. 1995. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*. Vol. 134. Elsevier, 789–811.

[140] Alejandro Peña, Humberto Sossa, and Agustín Gutiérrez. 2008. Causal knowledge and reasoning by cognitive maps: Pursuing a holistic approach. *Expert Systems with Applications* 35, 1-2 (2008), 2–18.

[141] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2013. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems* 26 (2013).

[142] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. 2014. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2035–2042.

[143] Drago Plecko and Elias Bareinboim. 2022. Causal fairness analysis. *arXiv preprint arXiv:2207.11385* (2022).

[144] Luigi Portinale. 1992. Verification of causal models using Petri nets. *International journal of intelligent systems* 7, 8 (1992), 715–742.

[145] Robert N Proctor. 2012. The history of the discovery of the cigarette–lung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco control* 21, 2 (2012), 87–91.

[146] Global Economic Prospects. 2020. Pandemic, Recession: The Global Economy in Crisis. *Washington DC: World Bank Group* (2020).

[147] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. Linked causal variational autoencoder for inferring paired spillover effects. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1679–1682.

[148] Manjusha Ravindranath, K Selçuk Candan, and Maria Luisa Sapino. 2020. M2NN: Rare event inference through multi-variate multi-scale attention. In *2020 IEEE International Conference on Smart Data Services (SMDS)*. IEEE, 53–62.

[149] Jack S Richards, Thangavelu U Arumugam, Linda Reiling, Julie Healer, Anthony N Hodder, Freya JI Fowkes, Nadia Cross, Christine Langer, Satoru Takeo, Alex D Uboldi, Jennifer K. Thompson, Paul R. Gilson, Ross L. Coppel, Peter M. Siba, Christopher L. King, Motomi Torii, Chetan E. Chitnis, David L. Narum, Ivo Mueller, Brendan S. Crabb, Alan F. Cowman, Takafumi Tsuboi, and James G. Beeson. 2013. Identification and prioritization of merozoite antigens as targets of protective human immunity to Plasmodium falciparum malaria for vaccine and biomarker development. *The Journal of Immunology* 191, 2 (2013), 795–809.

[150] Scott E Robinson, Warren S Eller, Melanie Gall, and Brian J Gerber. 2013. The core and periphery of emergency management networks. *Public Management Review* 15, 3 (2013), 344–362.

[151] Scott E Robinson, Brian J Gerber, Warren S Eller, and Melanie Gall. 2013. Emergency planning and disabled populations: assessing the FNSS approach. *International Journal of Mass Emergencies & Disasters* 31, 2 (2013), 315–329.

[152] DJ Rogers and SE Randolph. 2006. Climate change and vector-borne diseases. *Advances in parasitology* 62 (2006), 345–381.

[153] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688.

[154] Hanan Samet. 1984. The quadtree and related hierarchical data structures. *ACM Computing Surveys (CSUR)* 16, 2 (1984), 187–260.

[155] Hanan Samet. 1990. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley Longman Publishing Co., Inc., USA.

[156] Hanan Samet. 2005. *Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[157] Mohammad Reza Samsami, Mohammadhossein Bahari, Saber Salehkaleybar, and Alexandre Alahi. 2021. Causal imitative model for autonomous driving. *arXiv preprint arXiv:2112.03908* (2021).

[158] Claudio Schifanella, K. Selçuk Candan, and Maria Luisa Sapino. 2013. Multiresolution Tensor Decompositions with Mode Hierarchies. *ACM Trans. Knowl. Discov. Data* 8, 2 (2013), 10:1–10:38. https://doi.org/10.1145/2532169

[159] K. Selçuk Candan and Wen-Syan Li. 2000. Using Random Walks for Mining Web Document Associations. In *Knowledge Discovery and Data Mining. Current Issues and New Applications*, Takao Terano, Huan Liu, and Arbee L. P. Chen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 294–305.

[160] Timos Sellis, Nick Roussopoulos, and Christos Faloutsos. 1987. *The R+-Tree: A Dynamic Index for Multi-Dimensional Objects*. Technical Report.

[161] Paras Sheth, Ruocheng Guo, Lu Cheng, Huan Liu, and Kasim Selçuk Candan. 2023. Causal Disentanglement for Implicit Recommendations with Network Information. *ACM Trans. Knowl. Discov. Data* 17, 7 (2023), 94:1–94:18. https://doi.org/10.1145/3582435

[162] Paras Sheth, Ruocheng Guo, Kaize Ding, Lu Cheng, K Selçuk Candan, and Huan Liu. 2022. Causal disentanglement with network information for debiased recommendations. In *Similarity Search and Applications: 15th International Conference, SISAP 2022, Bologna, Italy, October 5–7, 2022, Proceedings*. Springer, 265–273.

[163] Paras Sheth, Ting Liu, Durmus Doner, Qi Deng, Yuhang Wei, Rebecca Muenich, John Sabo, K. Selçuk Candan, and Huan Liu. 2022. Causal Discovery for Feature Selection in Physical Process-Based Hydrological Systems. In *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, Shusaku Tsumoto, Yukio Ohsawa, Lei Chen, Dirk Van den Poel, Xiaohua Hu, Yoichi Motomura, Takuya Takagi, Lingfei Wu, Ying Xie, Akihiro Abe, and Vijay Raghavan (Eds.). IEEE, 5568–5577. https://doi.org/10.1109/BigData55660.2022.10020794

[164] Paras Sheth, Ahmadreza Mosallanezhad, Kaize Ding, Reepal Shah, John Sabo, Huan Liu, and K Selçuk Candan. 2023. STREAMS: Towards Spatio-Temporal Causal Discovery with Reinforcement Learning for Streamflow Rate Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4815–4821.

[165] Paras Sheth, Ahmadreza Mosallanezhad, Kaize Ding, Reepal Shah, John Sabo, Huan Liu, and K. Selçuk Candan. 2023. STREAMS: Towards Spatio-Temporal Causal Discovery with Reinforcement Learning for Streamflow Rate Prediction. In *Proceedings of the 32nd ACM International Conference*

on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 4815–4821. https://doi.org/10.1145/3583780.3614719

[166] Paras Sheth, Reepal Shah, John Sabo, K Selçuk Candan, and Huan Liu. 2022. STCD: A Spatio-Temporal Causal Discovery Framework for Hydrological Systems. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 5578–5583.

[167] Ilya Shpitser and Judea Pearl. 2012. Effects of treatment on the treated: Identification and generalization. *arXiv preprint arXiv:1205.2615* (2012).

[168] Edward H Simpson. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 13, 2 (1951), 238–241.

[169] Alan Siu and YC Richard Wong. 2004. Economic impact of SARS: The case of Hong Kong. *Asian Economic Papers* 3, 1 (2004), 62–83.

[170] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

[171] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, Vol. 3. SpringerOpen, 1–28.

[172] Jan Sprenger and Naftali Weinberger. 2021. Simpson's paradox. (2021).

[173] D Stampoulis, HG Damavandi, D Boscovic, and J Sabo. 2020. Using satellite remote sensing and machine learning techniques towards precipitation prediction and vegetation classification. *Journal of Environment Informatics* (2020), 1–45.

[174] Dimitrios Stampoulis, John T Reager, Cédric H David, Konstantinos M Andreadis, James S Famiglietti, Tom G Farr, Amy R Trangsrud, Ralph R Basilio, John L Sabo, Gregory B Osterman, Paul Lundgren, and Zhen Liu. 2019. Model-data fusion of hydrologic simulations and GRACE terrestrial water storage observations to estimate changes in water table depth. *Advances in Water Resources* 128 (2019), 13–27.

[175] Gui-Quan Sun, Marko Jusup, Zhen Jin, Yi Wang, and Zhen Wang. 2016. Pattern transitions in spatial epidemics: Mechanisms and emergent properties. *Physics of life reviews* 19 (2016), 43–73.

[176] Andrew J Tatem. 2009. The worldwide airline network and the dispersal of exotic species: 2007–2010. *Ecography* 32, 1 (2009), 94–102.

[177] Andrew J Tatem, Simon I Hay, and David J Rogers. 2006. Global traffic and disease vector dispersal. *Proceedings of the National Academy of Sciences* 103, 16 (2006), 6242–6247.

[178] Andrew J Tatem, David J Rogers, and Simon I Hay. 2006. Global transport networks and infectious disease spread. *Advances in parasitology* 62 (2006), 293–343.

[179] PanCommunity Project Team. 2023. NSF Smart and Connected Communities - PanCommunity Workshop. https://www.pancommunity.org/community-workshop.

[180] Johannes Textor, Juliane Hardt, and Sven Knüppel. 2011. DAGitty: a graphical tool for analyzing causal diagrams. *Epidemiology* 22, 5 (2011), 745.

[181] Jin Tian, Azaria Paz, and Judea Pearl. 1998. *Finding minimal d-separators*. Computer Science Department, University of California.

[182] Manoj Tiwaskar, Yash Garg, Xinsheng Li, K Selçuk Candan, and Maria Luisa Sapino. 2021. Selego: robust variate selection for accurate time series forecasting. *Data Mining and Knowledge Discovery* 35 (2021), 2141–2167.

[183] Edward C Tolman. 1948. Cognitive maps in rats and men. *Psychological review* 55, 4 (1948), 189.

[184] Hanghang Tong, Christos Faloutsos, and Yehuda Koren. 2007. Fast direction-aware proximity for graph mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 747–756.

[185] Thomas Verma and Judea Pearl. 1990. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*. Vol. 9. Elsevier, 69–76.

[186] Cécile Viboud, Lone Simonsen, and Gerardo Chowell. 2016. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. *Epidemics* 15 (2016), 27–37.

[187] Shu Wan, Reepal Shah, Qi Deng, John Sabo, Huan Liu, and K. Selçuk Candan. 2024. Spatiotemporal Causal Learning for Streamflow Forecasting. In *N-EWN Symposium (poster)*.

[188] Hang Wang, Zhenzhen Liu, Jun Zhu, Danjie Chen, and Fen Qin. 2022. Spatio-Temporal Extraction of Surface Waterbody and Its Response of Extreme Climate along the Upper Huaihe River. *Sustainability* 14, 6 (2022), 3223.

[189] Xiaolan Wang, K Selçuk Candan, and Maria Luisa Sapino. 2014. Leveraging metadata for identifying local, robust multi-variate temporal (RMT) features. In *2014 IEEE 30th International Conference on Data Engineering*. IEEE, 388–399.

[190] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. 2020. Causal Inference for Recommender Systems. In *Fourteenth ACM Conference on Recommender Systems*. 426–431.

[191] Fang Wei. 2010. TEDI: Efficient Shortest Path Query Answering on Graphs. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data* (Indianapolis, Indiana, USA) *(SIGMOD '10)*. Association for Computing Machinery, New York, NY, USA, 99–110. https://doi.org/10.1145/1807167.1807181

[192] Shaun Wilson and Norbert Ebert. 2013. Precarious work: Economic, sociological and political perspectives. , 263–278 pages.

[193] Sewall Wright. 1921. Correlation and causation. (1921).

[194] Joseph T Wu, Benjamin J Cowling, Eric HY Lau, Dennis KM Ip, Lai-Ming Ho, Thomas Tsang, Shuk-Kwan Chuang, Pak-Yin Leung, Su-Vui Lo, Shao-Haei Liu, and Steven Riley. 2010. School closure and mitigation of pandemic (H1N1) 2009, Hong Kong. *Emerging infectious diseases* 16, 3 (2010), 538.

[195] Yingjun Wu, Jia Yu, Yuanyuan Tian, Richard Sidle, and Ronald Barber. 2019. Designing succinct secondary indexing mechanism by exploiting column correlations. In *Proceedings of the 2019 International Conference on Management of Data*. 1223–1240.

[196] Yanghua Xiao, Wentao Wu, Jian Pei, Wei Wang, and Zhenying He. 2009. Efficiently Indexing Shortest Paths by Exploiting Symmetry in Graphs. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (Saint Petersburg, Russia) *(EDBT '09)*. Association for Computing Machinery, New York, NY, USA, 493–504. https://doi.org/10.1145/1516360.1516418

[197] Liangwei Yang, Zhiwei Liu, Yingtong Dou, Jing Ma, and Philip S Yu. 2021. Consisrec: Enhancing gnn for social recommendation via consistent neighbor aggregation. In *Proceedings of the 44th international ACM SIGIR conference on Research and development in information retrieval*. 2141–2145.

[198] Junko Yasuoka and Richard Levins. 2007. Impact of deforestation and agricultural development on anopheline ecology and malaria epidemiology. *The American journal of tropical medicine and hygiene* 76, 3 (2007), 450–460.

[199] Wenjia Zhang and Kexin Ning. 2023. Spatiotemporal heterogeneities in the causal effects of mobility intervention policies during the COVID-19 outbreak: A spatially interrupted time-series (SITS) analysis. *Annals of the American Association of Geographers* (2023), 1–23.

[200] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *Proceedings of the Web Conference 2021*. 2980–2991.

[201] Lei Zou, Lei Chen, and M. Tamer Özsu. 2009. Distance-Join: Pattern Match Query in a Large Graph Database. *Proc. VLDB Endow.* 2, 1 (aug 2009), 886–897. https://doi.org/10.14778/1687627.1687727

[202] Wlodek M Zuberek. 1991. Timed Petri nets definitions, properties, and applications. *Microelectronics Reliability* 31, 4 (1991), 627–644.